

L'intelligenza collettiva e i suoi limiti

Teoria dei giochi, meccanismo design e la governance dell'intelligenza artificiale

PREFAZIONE: IL QUINTO TERRITORIO

Il ciclo e la sua logica interna

Questo saggio conclude un ciclo di cinque monografie che hanno attraversato, con progressiva sistematicità e rigore crescente, il problema dei limiti dell'intelligenza — artificiale e umana insieme. Non si tratta di un ciclo costruito a posteriori, ritrovando retrospettivamente coerenza dove non era stata pianificata: il problema dei limiti era il filo conduttore esplicito fin dall'inizio, e la scelta dei cinque territori è stata dettata da una logica interna che vale la pena esplicitare prima di procedere, perché illumina il contributo specifico di questo quinto saggio rispetto ai predecessori.

Il primo contributo ha esplorato ciò che un sistema formale *può dimostrare*. Partendo dai teoremi di incompletezza di Gödel (1931) e dal problema dell'arresto di Turing (1936), ha mostrato che ogni sistema formale sufficientemente ricco da contenere l'aritmetica elementare è necessariamente incompleto — contiene proposizioni vere ma

indimostrabili dall'interno del sistema — oppure incoerente. Il teorema di Gödel non è un risultato tecnico sui sistemi assiomatici della matematica del primo Novecento: è una struttura che svela qualcosa di fondamentale sulla relazione tra verità e dimostrazione, tra ciò che è il caso e ciò che può essere certificato dall'interno di un sistema chiuso. La traduzione computazionale di questo risultato, nella forma del teorema di Rice, stabilisce che nessuna proprietà non banale dei sistemi computazionali è verificabile algebricamente in generale: la sicurezza di un programma arbitrario non può essere certificata da un altro programma con garanzia di terminazione. Questo limite non è tecnico ma strutturale; appartiene alla logica della computazione, non allo stato contingente delle nostre conoscenze tecniche.

Il secondo contributo ha esplorato ciò che una rappresentazione compressa può *catturare della struttura dell'informazione*. La geometria di Fisher e Rao, la sufficienza statistica, il principio di compressione di Kolmogorov: in ciascun caso è emerso che la rappresentazione non è mai neutrale rispetto all'informazione che veicola, e che esiste una geometria intrinseca — la varietà statistica equipaggiata con la metrica di Fisher — che misura la distanza tra distribuzioni in modo invariante rispetto alle scelte di parametrizzazione. I limiti della compressione nel senso di Shannon e i limiti della rappresentazione nel senso di Kolmogorov condividono la stessa struttura profonda: vi è sempre informazione che non può essere catturata senza perdita, e la quantità di questa perdita è misurabile con precisione.

Il terzo contributo ha esplorato ciò che un criterio di ottimizzazione può *specificare senza distorcere l'obiettivo originario*. La razionalità strumentale, il teorema di Arrow (1951), la legge di Goodhart: in ciascun caso il tema centrale era la distanza tra la misura e ciò che si intende misurare, tra il

criterio formale e il valore che si vuole promuovere. Il teorema di Arrow — forse il risultato più profondo dell'intera teoria della scelta sociale — mostra che nessuna funzione di aggregazione delle preferenze individuali può soddisfare simultaneamente unanimità, indipendenza dalle alternative irrilevanti e non-dittatura, quando le alternative sono più di due. La legge di Goodhart generalizza questa intuizione in modo operativo: quando una misura diventa un obiettivo, cessa di essere una buona misura, perché gli agenti ottimizzano la misura piuttosto che il fenomeno sottostante che la misura intendeva catturare. La razionalità strumentale non garantisce la razionalità degli obiettivi.

Il quarto contributo ha esplorato ciò che un agente può *essere in senso fenomenologico*, e se tale essere sia verificabile dall'esterno. Il problema difficile di Chalmers, il test di Turing e i suoi limiti, l'irriducibilità della prospettiva in prima persona: in ciascun caso è emersa l'impossibilità di colmare, dall'esterno e per via funzionale, il divario tra il comportamento osservabile e la vita interna — tra il *what it is like* e il *what it does*. La verificabilità della coscienza è strutturalmente impossibile per lo stesso motivo per cui la verificabilità della sicurezza computazionale è impossibile: si tratta di proprietà che non sono riducibili al comportamento osservabile su un insieme finito di situazioni.

Il quinto territorio: l'intelligenza collettiva

Questo saggio porta la stessa domanda di fondo — *dove finisce la portata di uno strumento?* — in un territorio qualitativamente diverso: non l'intelligenza singola, ma l'*intelligenza collettiva*. Come sistemi intelligenti molteplici, dotati di obiettivi che possono divergere e di informazione reciprocamente opaca, si coordinano verso esiti desiderabili? Quali strutture istituzionali rendono possibile questa coordinazione, e quali

tensioni le sono intrinseche indipendentemente dall'ingegnosità del progettista?

La domanda è diversa dalle precedenti non soltanto per il suo oggetto, ma per la sua struttura logica. Le quattro domande precedenti riguardavano ciò che è possibile *in linea di principio* per un sistema o un agente dato: i teoremi di impossibilità che le rispondevano valgono universalmente, indipendentemente dal contesto storico o tecnologico. La quinta domanda riguarda ciò che è possibile *nell'interazione*: e nell'interazione, il possibile dipende da chi interagisce, con quali incentivi, in quale contesto istituzionale, con quale distribuzione dell'informazione. I confini non sono fissi come quelli dei teoremi di Gödel: si spostano con il mutare delle condizioni.

Questo non rende la domanda meno importante — la rende più urgente, perché le sue risposte hanno conseguenze pratiche immediate. La governance dell'intelligenza artificiale, il coordinamento internazionale tra stati che sviluppano sistemi avanzati, la progettazione di meccanismi che allineino gli incentivi privati con il benessere collettivo: sono tutti aspetti della quinta domanda, che non trovano risposta nei teoremi di incompletezza o nella geometria dell'informazione, ma richiedono la teoria dei giochi, il meccanismo design e la teoria evolutiva.

Il problema della fondazione storica

Prima di addentrarsi nella materia, vale la pena collocare l'impresa intellettuale in un contesto storico più ampio. La questione della coordinazione collettiva tra agenti con interessi divergenti è una delle più antiche della filosofia politica. Da Hobbes, che nel *Leviathan* (1651) teorizzava il sovrano assoluto come unico rimedio al bellum omnium

contra omnes, a Locke e la sua fiducia nelle istituzioni contrattuali, fino a Rousseau e la sua volontà generale: il problema di come produrre ordine cooperativo a partire da agenti razionali ma potenzialmente egoisti ha attraversato i secoli senza trovare una risposta definitiva. La teoria dei giochi, inaugurata da von Neumann e Morgenstern nel 1944, non risolve questa tensione — la formalizza con precisione sufficiente da permetterne la manipolazione analitica. E la formalizzazione, come si vedrà, è già un passo di straordinaria importanza.

Ciò che la teoria dei giochi aggiunge all'intuizione filosofica è la possibilità di derivare, a partire da premesse esplicite e verificabili, conclusioni dimostrabili sui comportamenti di equilibrio. Quando Hobbes affermava che in assenza di sovrano la vita è "solitaria, povera, sgradevole, brutale e breve", stava esprimendo una tesi empirica difficilmente falsificabile. Quando la teoria dei giochi mostra che nel dilemma del prigioniero a gioco singolo la defezione è strategia dominante per entrambi i giocatori — indipendentemente da chi essi siano e da quali valori professino — sta derivando una conclusione logicamente necessaria date le strutture di payoff. La distanza tra i due livelli di analisi è la distanza tra la speculazione filosofica e la scienza formale.

Analogamente, Adam Smith intuiva nel 1776 che il mercato, attraverso la "mano invisibile", converte l'interesse privato in beneficio collettivo. Il meccanismo design dimostra esattamente *quando* questo accade e quando non accade: mostra i teoremi di impossibilità che delimitano i confini della mano invisibile, e costruisce le strutture di incentivi che possono estendere quei confini nel caso di informazione asimmetrica e interessi divergenti. Il teorema del benessere, il teorema di Myerson-Satterthwaite,

il principio di rivelazione: non sono oscurità matematiche — sono la risposta precisa alla domanda che Smith aveva posto con una metafora.

Una precisazione metodologica necessaria

Una precisazione si impone fin dall'inizio, perché la sua mancanza vizierebbe l'intera analisi successiva.

Sarebbe tentante presentare le difficoltà della coordinazione collettiva nello stesso registro delle quattro impossibilità precedenti: come confini logicamente necessari, matematicamente dimostrabili, che chiudono lo spazio del possibile con la stessa perentorietà di un teorema. Questo saggio resiste esplicitamente a quella tentazione, perché cederle sarebbe intellettualmente disonesto — e la disonestà intellettuale, in materia di governance, produce conseguenze pratiche gravi.

I teoremi di impossibilità dei saggi precedenti — Gödel, Arrow, Rice — valgono *universalmente*: sono veri in ogni contesto che soddisfa le loro condizioni, indipendentemente dal progresso tecnico o istituzionale. Le tensioni che emergeranno nell'analisi della coordinazione collettiva sono invece *strutturali* nel senso che emergono dalla geometria del problema, non dagli errori di chi lo affronta; sono *resistenti* nel senso che nessuna soluzione parziale le elimina completamente; ma sono *contingenti* rispetto al contesto in cui si manifestano. Ridurre l'incertezza tecnologica, costruire istituzioni più robuste, aumentare la trasparenza informativa: tutte queste azioni *attenuano* le tensioni strutturali della coordinazione, anche se non le eliminano. Nessuna di queste azioni può attenuare l'incompletezza di Gödel.

La distinzione non è accademica. Confondere i due registri — come fa molta letteratura sulla governance dell'AI che invoca impossibilità

Arrow-style per giustificare rassegnazione politica — produce confusione concettuale e, soprattutto, paralisi pratica ingiustificata. Se le difficoltà della governance fossero impossibilità matematiche universali, non vi sarebbe nulla da fare che non sia già nei limiti di ciò che la natura stessa del problema consente. Se invece sono tensioni strutturali contingenti, la progettazione istituzionale conserva un margine d'azione considerevole — a patto di capire con precisione la geometria di quelle tensioni, i loro punti di maggiore resistenza e i loro spazi di cedimento.

Ciò che questo saggio offre è dunque una *mappa* di quella geometria, tracciata con gli strumenti formali della teoria dei giochi e del meccanismo design. Una mappa non è il territorio: non risolve il problema della governance, non sostituisce la deliberazione politica, non decide quale compromesso privilegiare tra quelli che la struttura del problema rende necessari. Ma una carta accurata, in un territorio difficile e poco esplorato, ha un valore che non è piccolo — è la condizione per muoversi con consapevolezza invece di procedere alla cieca.

La matematica come lente

Vi è un'ultima questione metodologica da affrontare prima di procedere: il rapporto tra la matematica impiegata in questo saggio e la realtà empirica che essa intende illuminare.

La matematica è usata qui come *lente concettuale*, non come *modello* nel senso delle scienze esatte. Un modello aspira a descrivere il suo oggetto con sufficiente precisione da permettere previsioni quantitative verificabili. Una lente — nel senso in cui la teoria dei giochi è una lente sull'interazione strategica — è uno strumento che illumina certi aspetti strutturali del problema, chiarisce le relazioni di implicazione logica tra le

sue componenti, e rende esplicite le assunzioni necessarie a derivare certi risultati; ma non pretende di fornire una descrizione esaustiva della realtà, che conserva sempre una complessità irriducibile rispetto a qualsiasi formalizzazione.

Un agente economico reale non massimizza un'utilità attesa con la precisione di un elaboratore: ha capacità cognitive limitate, forma aspettative in modo euristico, è influenzato da norme sociali, stati emotivi e bias cognitivi ampiamente documentati dalla letteratura sperimentale. Eppure la teoria dell'utilità attesa illumina aspetti genuini del comportamento economico che senza di essa rimarrebbero opachi. La lente non è lo specchio: deforma selettivamente, amplificando ciò che intende analizzare e oscurando ciò che non rientra nel suo fuoco. La competenza analitica consiste nel sapere cosa la lente amplifica e cosa oscura — non nel credere che essa riproduca la realtà in modo fedele, né nel rifiutarla perché non lo fa.

Questa chiarezza metodologica è tanto più necessaria quanto più le conclusioni formali vengono trasportate nel dominio della politica pubblica. La traduzione dal risultato matematico alla raccomandazione di governance richiede sempre un passaggio interpretativo — un atto di giudizio su quale aspetto del modello cattura qualcosa di essenziale della realtà e quale invece è artefatto della semplificazione. Nelle pagine che seguono, questo passaggio interpretativo viene esplicitato ogni volta che è rilevante, invece di essere lasciato implicito come spesso accade nella letteratura applicativa.

La struttura del saggio

Il saggio è articolato in cinque parti principali, seguite da appendici tecniche che dettagliano le dimostrazioni più complesse.

La Parte Prima espone la struttura dell'interazione strategica: le tre forme in cui un gioco può essere rappresentato, i concetti fondamentali di equilibrio e le loro giustificazioni epistemiche, i principali risultati sulla struttura dei giochi a somma zero, gli equilibri correlati, il valore di Shapley, le raffinazioni dell'equilibrio e il prezzo dell'anarchia. La trattazione è formale ma motivata: ogni risultato matematico è presentato con la domanda che risponde e con le implicazioni che porta.

La Parte Seconda espone il meccanismo design: la sua rivoluzione concettuale, il principio di rivelazione, il meccanismo VCG e la sua caratterizzazione, il teorema di Myerson-Satterthwaite, l'asta ottimale, il meccanismo design multi-dimensionale e robusto, le dinamiche di impegno nel contesto dinamico, la teoria del matching. È la parte più tecnica del saggio, e la più ricca di implicazioni dirette per la progettazione istituzionale.

La Parte Terza espone la teoria dei giochi evolutivi: le dinamiche di replica, la legge di Fisher, la strategia evolutivamente stabile, le dinamiche stocastiche del processo di Moran, la struttura della rete e la condizione $b/c > k$, la selezione multi-livello e l'equazione di Price, la reciprocità indiretta, l'apprendimento multi-agente e la sua instabilità strutturale, il gradiente naturale come unificazione di evoluzione e apprendimento.

La Parte Quarta applica gli strumenti delle tre parti precedenti alla governance dell'intelligenza artificiale: la struttura dell'ecosistema, la selezione avversa nella certificazione, l'azzardo morale individuale e collettivo, la cattura del regolatore, i meccanismi di segnalazione credibile, la compute governance, i giochi globali e la coordinazione

internazionale, i principi di progettazione per meccanismi robusti, e la cattura evolutiva nel lungo periodo.

La Parte Quinta è sintetica: riflette su cosa la matematica rivela e cosa non può rivelare, nomina le quattro tensioni strutturali della governance con la precisione che meritano, applica il teorema del second-best come framework generale, e offre una prospettiva d'insieme sull'intero ciclo.

Le appendici tecniche raccolgono le dimostrazioni complete di cinque risultati fondamentali: l'esistenza dell'equilibrio di Nash attraverso il teorema di Kakutani, il teorema di Myerson-Satterthwaite per l'envelope theorem, la probabilità di fissazione nel processo di Moran, l'equivalenza tra dinamiche di replica e gradiente naturale di Fisher, e la struttura computazionale del politopo degli equilibri correlati.

LA STRUTTURA DELL'INTERAZIONE STRATEGICA

I.1 — Dalla decisione individuale all'interdipendenza

Il limite della teoria della decisione individuale

La teoria della decisione individuale — il framework della massimizzazione dell'utilità attesa sotto incertezza, che von Neumann e Morgenstern sistematizzarono nel 1944 e che Savage portò a maturità

con i suoi *Foundations of Statistics* del 1954 — è uno strumento analiticamente straordinario. La sua eleganza risiede nella separazione netta tra preferenze (rappresentate da una funzione di utilità) e credenze (rappresentate da una distribuzione di probabilità soggettiva), e nella regola di combinazione — l'aspettazione matematica — che produce la scelta ottimale. L'assiomatizzazione di Savage — completezza, transitività, indipendenza, continuità — riduce la razionalità a un insieme di condizioni di coerenza interna: chi le soddisfa si comporta *come se* massimizzasse un'utilità attesa, anche senza calcolare esplicitamente nulla.

La grandezza intellettuale di questa costruzione risiede, tra l'altro, nel fatto che il suo risultato chiave — la rappresentazione numerica delle preferenze soggettive — è un teorema, non un'assunzione. Se un agente soddisfa i sei assiomi di Savage, allora esiste necessariamente una funzione di utilità u e una misura di probabilità soggettiva π tali che l'agente, scegliendo tra azioni, si comporta come se stesse massimizzando l'aspettazione $\sum_s \pi(s) u(f(s))$ di ogni atto f . Il numero è conseguenza degli assiomi, non loro punto di partenza.

Ma questa architettura è intrinsecamente limitata quando applicata a situazioni in cui l'ambiente contiene altri agenti razionali le cui scelte dipendono, a loro volta, dalle proprie. La dipendenza è ricorsiva: la scelta ottimale di A dipende dalle aspettative su B , che dipendono dalle aspettative di B su A , che dipendono dalle aspettative di B su ciò che A si aspetta che B creda, e così via in una regressione potenzialmente infinita. Questa ricorsività — talvolta chiamata *interdipendenza delle aspettative* — non è un caso patologico: è la struttura normale di qualsiasi interazione strategica, dai mercati alle negoziazioni diplomatiche, dalla concorrenza tra imprese al coordinamento tra sistemi AI multi-agente.

Il collasso del reinforcement learning in ambienti multi-agente

Anche nel reinforcement learning nella sua formulazione standard come processo decisionale di Markov (MDP), questa ricorsività è assente per ipotesi. Il teorema di Bellman stabilisce che la funzione valore ottimale V^* soddisfa:

$$V^*(s) = \max_a [R(s,a) + \gamma \cdot \sum_{s'} P(s'|s,a) \cdot V^*(s')]$$

dove $P(s'|s,a)$ è la funzione di transizione dell'ambiente — esogena, stabile, non reattiva alle strategie dell'agente. La struttura dell'equazione è quella di un'ottimizzazione su un singolo agente in un ambiente fisso. La ricorsività è temporale (il valore futuro entra nel calcolo presente) ma non strategica: nessun altro agente modifica l'ambiente in risposta alle nostre azioni.

Non appena P dipende dalle azioni di un secondo agente che persegue obiettivi propri, la struttura si dissolve. Non esiste più una funzione valore ottimale in senso assoluto, perché il valore di un'azione dipende dall'azione simultanea o futura dell'agente avversario, che dipende a sua volta dal comportamento atteso del primo. Il problema si ridefinisce come ricerca di un *punto fisso nello spazio delle politiche congiunte* — ed è precisamente qui che nasce la teoria dei giochi.

Questa osservazione non è accademica. Il training di sistemi AI in ambienti multi-agente — nei mercati di raccomandazione dove più modelli competono per l'attenzione degli utenti, nei sistemi di trading algoritmico dove più agenti operano sullo stesso mercato, nei contesti in cui modelli di linguaggio interagiscono tra loro o con esseri umani in modi che modificano i dati su cui i modelli futuri verranno addestrati —

affronta esattamente questo problema. Ignorarlo, trattando l'ambiente come stazionario mentre altri agenti si adattano, porta a instabilità, oscillazioni e comportamenti emergenti non intenzionali che nessun singolo progettista aveva previsto.

La fondazione storica del campo

La fondazione sistematica della teoria dei giochi si deve a John von Neumann e Oskar Morgenstern, che nella *Theory of Games and Economic Behavior* del 1944 posero le basi matematiche dell'analisi delle situazioni di conflitto e cooperazione tra agenti razionali. Von Neumann aveva dimostrato il teorema del minimax già nel 1928 — il risultato fondamentale per i giochi a somma zero a due giocatori — ma la generalizzazione a giochi con più partecipanti e interessi non diametralmente opposti richiedeva un nuovo framework concettuale. La collaborazione con Morgenstern produsse un'opera che era allo stesso tempo fondazione assiomatica della teoria dell'utilità e analisi dell'interazione strategica: un'impresa intellettuale di portata raramente eguagliata nella storia della scienza sociale.

Il contributo di John Nash del 1950-51 trasformò il campo in modo definitivo. In due lavori straordinariamente concisi — "Equilibrium points in n -person games" (1950, tre pagine) e "Non-cooperative games" (1951) — Nash definì il concetto di equilibrio che porta il suo nome e ne dimostrò l'esistenza per qualsiasi gioco finito in strategie miste, usando il teorema del punto fisso di Kakutani. L'intuizione di Nash era di disarmante semplicità una volta formulata: in un equilibrio, nessun giocatore ha incentivo a deviare unilateralmente dalla propria strategia, dato ciò che gli altri stanno facendo. La difficoltà era dimostrare che un tale punto fisso esiste sempre, anche in giochi con strutture molto

generali. Il Nobel per l'economia del 1994, assegnato congiuntamente a Nash, Harsanyi e Selten, ha sancito la centralità di questi contributi.

Il contributo di Harsanyi, premiato nello stesso anno, riguardava la modellizzazione dell'informazione asimmetrica attraverso i tipi privati — l'idea che l'incertezza su chi sono gli avversari possa essere rappresentata come incertezza sulla loro "natura" estratta da una distribuzione comune. Selten, da parte sua, ha sviluppato le raffinazioni dell'equilibrio che eliminano le soluzioni appoggiate su minacce non credibili. Insieme, i tre filoni hanno costruito la teoria dei giochi moderna nella sua forma attuale.

I.2 — Le tre forme di un gioco

La forma normale

La *forma normale* (o strategica) descrive una situazione in cui i giocatori scelgono simultaneamente le loro azioni, senza osservare le scelte altrui prima di agire. Formalmente, un gioco in forma normale è una terna:

$$G = (N, \{A_i\}_{i \in N}, \{u_i\}_{i \in N})$$

dove $N = \{1, \dots, n\}$ è l'insieme dei giocatori, A_i è l'insieme delle azioni disponibili al giocatore i , e $u_i : \times_j A_j \rightarrow \mathbb{R}$ è la funzione di utilità. Il profilo di azioni è $a \in A = \times_i A_i$, e si usa la notazione a_{-i} per le azioni di tutti i giocatori tranne i .

Una *strategia mista* $\sigma_i \in \Delta(A_i)$ è una distribuzione di probabilità sull'insieme delle azioni pure. L'utilità attesa sotto il profilo di strategie miste σ è:

$$U_i(\sigma) = \sum_{a \in A} [\prod_j \sigma_j(a_j)] \cdot u_i(a)$$

La motivazione per l'introduzione delle strategie miste va oltre la necessità tecnica di garantire l'esistenza dell'equilibrio. Una strategia mista rappresenta genuinamente l'incertezza che ciascun giocatore mantiene sulle scelte degli avversari: se un giocatore non sa con certezza cosa farà l'avversario, il suo processo decisionale è equivalente, dal punto di vista esterno, a quello di qualcuno che randomizza. L'indistinguibilità tra incertezza sull'avversario e randomizzazione propria è, in fondo, la ragione profonda per cui le strategie miste sono il contesto naturale dell'analisi di equilibrio.

Un gioco è *a somma zero* se $\sum_i u_i(a) = 0$ per ogni a ; *a somma variabile* altrimenti. Questa distinzione è fondamentale: nei giochi a somma zero il guadagno di uno è la perdita dell'altro, e non esiste possibilità di guadagno cooperativo; nei giochi a somma variabile esistono potenziali guadagni dalla cooperazione che nessun giocatore può realizzare unilateralmente. La governance dell'AI è un gioco a somma variabile: esistono configurazioni dell'ecosistema in cui tutti gli attori stanno meglio rispetto all'equilibrio non cooperativo, ma raggiungerle richiede coordinazione che i soli incentivi di mercato non garantiscono.

La forma estesa

La *forma estesa* rappresenta i giochi sequenziali attraverso un albero orientato. I nodi rappresentano i punti di scelta, gli archi le azioni disponibili, i nodi terminali i payoff. La struttura informativa è catturata dagli *insiemi informativi*: raggruppamenti di nodi che un giocatore non sa distinguere al momento della scelta — perché non ha osservato le mosse precedenti degli altri, o perché non conosce la propria posizione nell'albero. Quando ogni insieme informativo è un singoletto, il gioco è *a*

informazione perfetta: scacchi, go, la maggior parte dei giochi di strategia sequenziali. Quando esistono insiemi informativi non singolari, il gioco è *a informazione imperfetta*: poker, la maggior parte delle negoziazioni reali, i giochi di governance tra attori con asimmetrie informative strutturali.

Una *strategia* per il giocatore i è un piano completo di azione: una funzione che specifica quale azione intraprendere in *ogni* insieme informativo, inclusi quelli che non verrebbero mai raggiunti se tutti i giocatori seguissero le strategie di equilibrio. Questa completezza è cruciale: è ciò che rende possibile valutare se le minacce sono credibili — una minaccia è credibile solo se il giocatore che la profferisce avrebbe effettivamente convenienza a eseguirla se il nodo rilevante venisse raggiunto. La completezza della strategia estende così il concetto di razionalità oltre le situazioni di equilibrio, imponendo coerenza anche nei controfattuali.

Il **teorema di Kuhn (1953)** garantisce che, per giochi con *ricordo perfetto* — ogni giocatore ricorda le proprie azioni passate e le informazioni osservate — le strategie miste e le strategie comportamentali sono equivalenti: producono la stessa distribuzione sugli esiti. Una strategia comportamentale specifica, per ogni insieme informativo, una distribuzione di probabilità sulle azioni disponibili in quell'insieme; una strategia mista specifica una distribuzione sull'insieme delle strategie pure complete. La loro equivalenza semplifica enormemente l'analisi.

La forma cooperativa

La *forma cooperativa* (o coalizionale) astrae dall'interazione specifica e descrive i giochi attraverso la *funzione caratteristica* $v : 2^N \rightarrow \mathbb{R}$, che assegna a ogni coalizione $S \subseteq N$ il massimo valore che essa può garantire ai propri membri, indipendentemente dal comportamento dei

giocatori esterni. La funzione caratteristica è *super-additiva* se $v(S \cup T) \geq v(S) + v(T)$ per coalizioni disgiunte: la cooperazione non è mai controproducente e tipicamente genera valore aggiuntivo.

La forma cooperativa è il terreno naturale delle soluzioni basate sull'equità: il nucleo (le allocazioni che nessuna coalizione ha incentivo a rifiutare), il valore di Shapley (l'unica soluzione che soddisfa certi assiomi di equità), il nucleo di contrattazione. Queste soluzioni non descrivono il *processo* di negoziazione ma l'*esito* che risulterebbe da una negoziazione tra parti con piena capacità di impegno vincolante — un'astrazione utile per pensare a come dovrebbero essere distribuite le responsabilità in un ecosistema AI cooperativo, anche se il mondo reale raramente soddisfa le ipotesi di impegno.

I.3 — L'equilibrio di Nash: definizione, esistenza e interpretazioni

Definizione

Un profilo di strategie $\sigma^* = (\sigma_1^*, \dots, \sigma_n^*)$ è un *equilibrio di Nash* se nessun giocatore ha incentivo a deviare unilateralmente:

$$\forall i \in N, \forall \sigma_{-i}' \in \Delta(A_{-i}): U_i(\sigma_i^*, \sigma_{-i}') \geq U_i(\sigma_i^*, \sigma_{-i}^*)$$

dove $U_i(\sigma) = \sum_{a \in A} [\prod_j \sigma_j(a_j)] \cdot u_i(a)$ è l'utilità attesa sotto il profilo di strategie miste σ .

In strategie miste, la condizione di equilibrio ammette una caratterizzazione equivalente e illuminante. Poiché U_i è lineare in σ_i , la

best response di i a σ_{-i} è generalmente un insieme convesso di strategie miste che assegnano probabilità positiva soltanto alle azioni pure che massimizzano U_i . Dunque σ^* è un equilibrio se e solo se: per ogni giocatore i , ogni azione a_i nel supporto di σ_i^* è una best response a σ_{-i}^* . Ne segue la *condizione di indifferenza*: il giocatore è indifferente tra tutte le azioni nel supporto della propria strategia mista. Questa condizione vincola in modo forte la struttura degli equilibri misti — un giocatore randomizza tra azioni diverse soltanto se è indifferente, e questa indifferenza non è garantita dalla propria strategia ma dalle strategie degli avversari. È un punto sottile ma cruciale: le proprie probabilità di randomizzazione sono determinante dall'ottimizzazione degli avversari, non dalla propria.

Esistenza: il teorema di Nash via Kakutani

La dimostrazione dell'esistenza di almeno un equilibrio in strategie miste per ogni gioco finito procede in tre passi logici che vale la pena articolare con cura.

Definiamo la *corrispondenza di best response* di ogni giocatore:

$$BR_i(\sigma_{-i}) = \operatorname{argmax}_{\{\sigma_i \in \Delta(A_i)\}} U_i(\sigma_i, \sigma_{-i})$$

e la corrispondenza di best response congiunta:

$$BR(\sigma) = \times_i BR_i(\sigma_{-i})$$

Un equilibrio di Nash è esattamente un punto fisso di BR : $\sigma^* \in BR(\sigma^*)$.

Il problema si riduce a dimostrare che BR ha un punto fisso.

Il **teorema del punto fisso di Kakutani (1941)** garantisce l'esistenza del punto fisso sotto quattro condizioni, tutte soddisfatte nel nostro contesto:

Lo spazio $\Delta = \times_i \Delta_i$ è compatto (prodotto di semplici standard, chiusi e limitati in $\mathbb{R}^{|A_i|}$) e convesso (combinazioni convesse di strategie miste sono strategie miste). Questo verifica la prima condizione.

$BR(\sigma)$ è non-vuota per ogni σ : segue dal teorema del massimo, in quanto U_i è continua su Δ_i compatto, quindi il massimo è raggiunto. Seconda condizione verificata.

$BR(\sigma)$ è convessa per ogni σ : segue dalla linearità di U_i in σ_i . Se σ_i e σ_i' sono entrambe best response a σ_{-i} , allora per qualsiasi $\lambda \in [0,1]$:

$$\begin{aligned} U_i(\lambda\sigma_i + (1-\lambda)\sigma_i', \sigma_{-i}) &= \lambda U_i(\sigma_i, \sigma_{-i}) + (1-\lambda)U_i(\sigma_i', \sigma_{-i}) \\ &= \max_{\sigma_i''} U_i(\sigma_i'', \sigma_{-i}) \end{aligned}$$

Dunque anche la combinazione convessa è una best response. Terza condizione verificata.

Il grafo di BR è chiuso: se $\sigma^k \rightarrow \sigma$ e $\tau^k \rightarrow \tau$ con $\tau^k \in BR(\sigma^k)$, allora al limite $\tau \in BR(\sigma)$, per continuità di U_i . Quarta condizione verificata.

Il teorema di Kakutani garantisce l'esistenza di $\sigma^* \in BR(\sigma^*)$, che è per definizione un equilibrio di Nash. $\square$

Vale la pena sottolineare cosa questo teorema *non* garantisce: non garantisce l'unicità, non garantisce l'efficienza, non garantisce che l'equilibrio sia selezionabile in modo naturale tra i molteplici candidati. La molteplicità degli equilibri è la norma nei giochi con struttura non triviale — e la selezione tra equilibri è un problema distinto che occuperà gran parte dell'analisi successiva.

Le interpretazioni dell'equilibrio

Il dibattito sulla giustificazione epistemica dell'equilibrio di Nash è uno dei più ricchi e irrisolti della teoria economica, con implicazioni dirette per la progettazione di meccanismi di governance.

La *lettura introspettiva* vede l'equilibrio come la previsione cui agenti razionali convergerebbero attraverso ragionamento puro. Se è *common knowledge* che tutti i giocatori sono razionali — che ciascuno sa che tutti gli altri sono razionali, che ciascuno sa che tutti sanno che tutti sono razionali, e così via a tutti i livelli — allora ciascuno dovrebbe essere in grado di calcolare la strategia dell'altro, e quindi la propria best response. Il problema è che questo ragionamento non porta univocamente all'equilibrio in presenza di molteplici equilibri: per selezionarne uno specifico si devono fare assunzioni aggiuntive sulla salienza, sulla storia delle interazioni, sui meccanismi focali — che vanno oltre la razionalità individuale.

Un problema ancora più profondo è messo in luce da Bernheim (1984) e Pearce (1984) con il concetto di *razionalizzabilità*: la common knowledge della razionalità implica soltanto che ciascun giocatore adotti una strategia razionalizzabile — una strategia che è best response a qualche distribuzione di credenze sulle strategie degli avversari che sia essa stessa razionalizzabile. L'insieme delle strategie razionalizzabili è in genere più ampio dell'insieme degli equilibri di Nash: la common knowledge della razionalità non è sufficiente a selezionare l'equilibrio. Si tratta di un risultato scomodo: l'unico concetto epistemico pienamente giustificabile — la common knowledge della razionalità — non è sufficiente a giustificare l'equilibrio di Nash.

La *lettura epistemica*, sviluppata con rigore da Aumann e Brandenburger (1995), chiarisce esattamente quali condizioni sulla conoscenza reciproca dei giocatori siano necessarie e sufficienti per l'equilibrio. Per giochi a due

giocatori: razionalità di entrambi, mutua conoscenza delle credenze, e mutua conoscenza della razionalità. La condizione cruciale — che ciascuno *conosca esattamente le credenze dell'altro su se stesso* — è molto più forte della sola razionalità e generalmente non verificabile in contesti reali.

La *lettura evolutiva* — fondamento della Parte Terza di questo saggio — vede l'equilibrio non come risultato di un calcolo deliberato ma come stato stazionario di processi di aggiustamento adattivo in cui le strategie più performanti si diffondono a spese delle meno performanti. Questa interpretazione offre una giustificazione robusta che non richiede le forti assunzioni cognitive delle letture precedenti. Il suo costo è che non tutti gli equilibri di Nash sono dinamicamente stabili: solo certi equilibri sono attrattori del processo evolutivo, e la teoria deve specificare quale meccanismo di aggiustamento si considera.

I.4 — Il teorema del minimax e la struttura dei giochi a somma zero

Formulazione

In un gioco a somma zero con due giocatori — dove $u_1(a) + u_2(a) = 0$ per ogni a , dunque $u_2 = -u_1$ — il giocatore I massimizza ciò che il giocatore II minimizza. In strategie pure, i valori di sicurezza soddisfano in generale la disuguaglianza:

$$\max_{a_1} \min_{a_2} u_1(a_1, a_2) \leq \min_{a_2} \max_{a_1} u_1(a_1, a_2)$$

La disuguaglianza riflette il fatto che annunciare la propria strategia prima dell'avversario è uno svantaggio: il giocatore che si muove per primo può essere sfruttato da chi risponde ottimalmente. Questo vale in strategie pure. In strategie miste si ha il risultato fondamentale.

Il **teorema del minimax di von Neumann (1928)** afferma che in strategie miste i due valori coincidono:

$$\max_{\sigma_1} \min_{\sigma_2} U_1(\sigma_1, \sigma_2) = \min_{\sigma_2} \max_{\sigma_1} U_1(\sigma_1, \sigma_2) = V^*$$

dove V^* è il *valore del gioco*.

Dimostrazione via dualità lineare

La dimostrazione più elegante sfrutta la **dualità forte della programmazione lineare**. Sia M la matrice dei payoff del giocatore I, di dimensioni $m \times n$. Il problema di massimizzazione del giocatore I — che sceglie $\sigma_1 \in \Delta^m$ per massimizzare il minimo payoff garantito — può essere scritto come:

$$\max_{\sigma_1, v} v \text{ soggetto a: } M^T \sigma_1 \geq v \cdot \mathbf{1}_n, \sigma_1 \geq 0, \sum_k (\sigma_1)_k = 1$$

Il problema di minimizzazione del giocatore II — che sceglie $\sigma_2 \in \Delta^n$ per minimizzare il massimo payoff concesso — è:

$$\min_{\sigma_2, w} w \text{ soggetto a: } M \sigma_2 \leq w \cdot \mathbf{1}_m, \sigma_2 \geq 0, \sum_j (\sigma_2)_j = 1$$

Questi due programmi lineari sono duali l'uno dell'altro nel senso della programmazione lineare. Per la dualità forte — che vale perché entrambi i problemi sono ammissibili e limitati — i valori ottimali coincidono: il

massimo del primale uguaglia il minimo del duale. Questa coincidenza è esattamente l'uguaglianza delle strategie minimax. $\square$

Proprietà strutturali e implicazioni

Il valore V^* è una quantità *oggettiva* del gioco: non dipende dall'identità di chi annuncia per primo la propria strategia, perché in strategie miste l'ordine di annuncio non è rilevante. Le strategie minimax σ_1^* e σ_2^* sono ottimalmente robuste alla mancanza di conoscenza dell'avversario: σ_1^* garantisce al giocatore I almeno V^* indipendentemente dalla strategia del giocatore II, e ogni deviazione da σ_1^* può essere sfruttata dall'avversario per portare il payoff di I al di sotto di V^* .

Questa robustezza è precisamente ciò che rende le strategie minimax interessanti per la governance: in ambienti con avversari strategicamente capaci e intenzioni opache — sistemi AI potenzialmente disallineati, attori statali con obiettivi competitivi, industrie con incentivi a manipolare i regolatori — le strategie che garantiscono il peggio-caso hanno un valore assicurativo che le strategie ottimizzate per lo scenario atteso non possiedono.

Il minimax ha anche una connessione profonda con la teoria della complessità algoritmica, che vale menzionare. Per la classe dei giochi a somma zero ripetuti, la strategia minimax ottimale può essere computata in tempo polinomiale come soluzione di un programma lineare. Questo è in netto contrasto con il calcolo degli equilibri di Nash nei giochi generali, che è PPAD-completo — un risultato di Daskalakis, Goldberg e Papadimitriou (2006) che suggerisce l'intrinseca difficoltà computazionale dell'equilibrio in giochi non a somma zero.

I.5 — L'equilibrio correlato: fondamento epistemico superiore

Definizione e struttura

L'*equilibrio correlato*, introdotto da Robert Aumann nel 1974, è uno dei contributi più profondi e sottovalutati della teoria dei giochi moderna. Mentre l'equilibrio di Nash assume che i giocatori randomizzino in modo *indipendente*, l'equilibrio correlato ammette che le loro azioni siano correlate attraverso un segnale comune — un dispositivo di correlazione a cui tutti i giocatori hanno accesso, ma ciascuno osserva soltanto il segnale che gli è diretto.

Formalmente, una distribuzione μ sull'insieme dei profili di azioni $A = \times_i A_i$ è un equilibrio correlato se, per ogni giocatore i e ogni funzione di deviazione $\tau_i : A_i \rightarrow A_i$:

$$\sum_{a \in A} \mu(a) u_i(a) \geq \sum_{a \in A} \mu(a) u_i(\tau_i(a_i), a_{-i})$$

La lettura è la seguente: un dispositivo esterno campiona un profilo d'azioni a dalla distribuzione μ e raccomanda privatamente a ogni giocatore i la propria componente a_i . Il profilo è un equilibrio correlato se ciascun giocatore preferisce seguire la propria raccomandazione, supponendo che gli altri la seguano — e questa preferenza deve valere per ogni possibile realizzazione a_i che il giocatore potrebbe aver ricevuto come raccomandazione.

Il semaforo è l'esempio prototipico: entrambi gli automobilisti osservano lo stesso segnale (il colore del semaforo), ma ricevono raccomandazioni diverse (verde per uno, rosso per l'altro). Ciascuno ha incentivo a seguire la propria raccomandazione se si aspetta che l'altro faccia altrettanto. Il

dispositivo non è un arbitro che impone le azioni: è uno strumento di coordinazione delle aspettative che rende ciascuna raccomandazione auto-applicante.

Tre proprietà fondamentali

La prima proprietà riguarda la struttura geometrica. L'insieme degli equilibri correlati di un gioco finito è un *politopo convesso*, definito da un sistema di disuguaglianze lineari. Questo ha un'implicazione computazionale cruciale: ottimizzare una funzione obiettivo lineare — come il benessere sociale — sugli equilibri correlati è un problema di programmazione lineare, risolvibile in tempo polinomiale. La computabilità degli equilibri correlati li rende strumenti pratici di coordinazione, non solo oggetti teorici.

La seconda proprietà è il contenimento: ogni equilibrio di Nash in strategie miste è un equilibrio correlato, perché le distribuzioni prodotte $\mu = \times_i \sigma_i$ soddisfano le condizioni di equilibrio correlato con correlazione nulla. L'insieme degli equilibri correlati è quindi più ricco di quello degli equilibri di Nash, e può contenere distribuzioni che producono payoff superiori a qualsiasi equilibrio di Nash. Questo rende possibile la coordinazione verso esiti migliori senza richiedere l'allineamento esplicito delle strategie.

La terza proprietà è il fondamento epistemico. Aumann (1987) dimostrò che ogni stato del mondo in cui i giocatori sono *razionali in senso bayesiano* e le loro credenze derivano da un *prior comune* genera necessariamente un equilibrio correlato. La condizione del prior comune — che i giocatori abbiano credenze ex ante identiche prima di osservare i propri segnali privati — è molto più debole della common knowledge delle strategie richiesta per giustificare l'equilibrio di Nash. In questo senso preciso,

L'equilibrio correlato è la nozione di soluzione che emerge *naturalmente* dalla razionalità bayesiana, senza richiedere coordinazione sui comportamenti. Questo risultato inverte la prospettiva tradizionale: invece di chiedersi se gli agenti razionali giocherebbero un equilibrio correlato, mostra che se sono razionali e condividono un prior, giocano *necessariamente* un equilibrio correlato.

Un esempio illustrativo: il gioco del pollo

Nel gioco del "pollo" — due guidatori si avvicinano su una corsia stretta, ciascuno può deviare (D) o andare dritto (S) — la matrice dei payoff è la seguente. Se entrambi deviano, il payoff è (0,0). Se il giocatore 1 devia e il 2 va dritto, il payoff è (1,2). Se il giocatore 1 va dritto e il 2 devia, è (2,1). Se entrambi vanno dritto, è (-1,-1).

Gli equilibri di Nash puri sono (D,S) e (S,D) con payoff (1,2) e (2,1) rispettivamente. L'equilibrio misto ha entrambi che vanno dritto con probabilità $1/3$, producendo payoff attesi $(2/3, 2/3)$ — inferiore a qualsiasi equilibrio puro per almeno uno dei giocatori.

L'equilibrio correlato ottimo assegna probabilità $1/2$ a (D,S) e $1/2$ a (S,D), producendo payoff attesi $(3/2, 3/2)$ — superiore a qualsiasi equilibrio di Nash per *entrambi* i giocatori. Un semaforo che dice casualmente a uno di deviare e all'altro di andare dritto realizza questo equilibrio senza richiedere coordinazione esplicita sulle azioni. Il guadagno cooperativo rispetto all'equilibrio misto è $(3/2 - 2/3) = 5/6$ per ciascun giocatore — un aumento del 125%. La correlazione vale oro.

I.6 — Il valore di Shapley: equità assiomatica nell'allocazione cooperativa

Assiomi e caratterizzazione

Nella teoria cooperativa, la domanda fondamentale è come distribuire il valore creato dalla grande coalizione N tra i suoi membri, in modo da riflettere equamente il contributo di ciascuno. Lloyd Shapley (1953) propose quattro assiomi che catturano l'idea intuitiva di equità e dimostrò che esiste un'unica soluzione che li soddisfa tutti.

Gli assiomi sono: efficienza ($\sum_i \varphi_i(v) = v(N)$, l'intero valore viene distribuito), simmetria (se i e j sono giocatori equivalenti nel senso che $v(S \cup \{i\}) = v(S \cup \{j\})$ per ogni coalizione S che non li contiene, allora $\varphi_i(v) = \varphi_j(v)$), giocatore fittizio (se i non contribuisce a nessuna coalizione nel senso che $v(S \cup \{i\}) = v(S)$ per ogni S , allora $\varphi_i(v) = 0$), e addittività ($\varphi(v+w) = \varphi(v) + \varphi(w)$ per qualsiasi coppia di giochi v, w sullo stesso insieme di giocatori N).

La soluzione unica — il valore di Shapley — è:

$$\varphi_i(v) = \sum_{\{S \subseteq N \mid i \in S\}} \left[\frac{|S|!(|N|-|S|-1)!}{|N|!} \right] \cdot [v(S \cup \{i\}) - v(S)]$$

Il fattore combinatorio $|S|!(|N|-|S|-1)!/|N|!$ è la probabilità che, in un ordinamento casuale uniforme dei giocatori, la coalizione che si trova davanti a i sia esattamente S . Il valore di Shapley è dunque il *contributo marginale atteso* di ciascun giocatore, mediato su tutti gli ordinamenti possibili di arrivo nella grande coalizione.

Interpretazione e applicazioni alla governance

L'interpretazione è trasparente: il valore di Shapley distribuisce il valore proporzionalmente al contributo marginale medio. Un giocatore il cui contributo marginale è elevato in molte coalizioni riceve una quota ampia; un giocatore con contributo marginale basso riceve poco. Questo è intuitivamente equo, e l'assiomatizzazione formalizza in modo preciso in quale senso lo sia.

Il valore di Shapley è rilevante per la governance dell'AI in almeno tre contesti distinti.

Il primo riguarda la **distribuzione dei benefici e dei costi** tra gli stakeholder di un sistema fondazionale: se diverse organizzazioni collaborano allo sviluppo di un modello di base, il valore di Shapley fornisce una base assiomaticamente giustificata per la distribuzione dei profitti e dei costi di sicurezza, proporzionalmente al contributo marginale di ciascuna.

Il secondo riguarda la **valutazione del contributo dei dati**: il *data Shapley* (Ghorbani e Zou, 2019) applica la formula ai dataset di addestramento, trattando ciascun dato come un "giocatore" che contribuisce alla performance del modello. Questo ha implicazioni immediate per la governance dei dati — inclusa la determinazione della responsabilità legale in caso di danni causati da comportamenti appresi su dati specifici — e per i diritti di proprietà intellettuale nel contesto del training su dati prodotti da terzi. La computazione del data Shapley è in genere costosa ($O(n!)$ nel caso peggiore), ma esistono approssimazioni efficienti basate sul campionamento casuale delle permutazioni.

Il terzo riguarda la **distribuzione delle responsabilità regolamentari** in sistemi multi-livello: se lo sviluppatore di un modello di base, il deployer di un'applicazione e l'utente finale contribuiscono ciascuno alla

catena causale di un danno, il valore di Shapley fornisce una base per distribuire la responsabilità legale proporzionalmente al contributo causale di ciascun attore — una questione che nessun sistema giuridico vigente affronta con strumenti formalmente rigorosi.

I limiti del valore di Shapley meritano altrettanta attenzione. La soluzione non è necessariamente nell'*insieme del nucleo* — l'insieme delle allocazioni che nessuna coalizione ha incentivo a rifiutare. Quando il nucleo è vuoto, nessuna allocazione è stabile e il valore di Shapley fornisce una soluzione equa ma non stabile. Nei giochi di governance dell'AI, dove le coalizioni tra organizzazioni sono instabili e le minacce di uscita sono credibili, questa distinzione è rilevante e non può essere ignorata nel design istituzionale.

I.7 — Raffinazioni dell'equilibrio: credibilità e coerenza

Il problema delle minacce non credibili

Il teorema di Nash garantisce l'esistenza di almeno un equilibrio, ma non l'unicità: la molteplicità è la norma nei giochi con struttura non triviale. Nei giochi sequenziali, la molteplicità è aggravata dalla presenza di equilibri che dipendono da *minacce non credibili*: strategie che specificano comportamenti ai nodi fuori equilibrio — nodi che non verrebbero mai raggiunti se i giocatori seguissero le proprie strategie di equilibrio — che non sarebbero razionalmente eseguiti se quei nodi venissero effettivamente raggiunti.

Il **gioco dell'entrata nel mercato** illustra il problema in forma elementare. Un entrante E decide se entrare o restare fuori; il monopolista M risponde con accomodamento A o guerra dei prezzi G. I payoff sono:

$$u_E(\text{Entra}, A) = 2, u_M(\text{Entra}, A) = 1 \quad u_E(\text{Entra}, G) = -1, u_M(\text{Entra}, G) = -1 \\ u_E(\text{Fuori}) = 0, u_M(\text{Fuori}) = 3$$

Esistono due equilibri di Nash: (Entra, A) con payoff (2,1) e (Fuori, G) con payoff (0,3). Il secondo è sorretto dalla minaccia di M di fare la guerra dei prezzi se E entra. Ma questa minaccia è vuota: se E entra, M affronta la scelta tra A (payoff 1) e G (payoff -1), e preferisce accomodare. La minaccia di guerra non sarebbe eseguita — è strategicamente irrazionale ex post.

L'equilibrio sottogame-perfetto di Selten

L'*equilibrio sottogame-perfetto* (Selten, 1965) elimina questi equilibri richiedendo che le strategie costituiscano un equilibrio di Nash in ogni *sottogame* — ogni sotto-albero del gioco che inizia da un nodo singolo che è al contempo l'unico elemento del proprio insieme informativo. Il calcolo si effettua attraverso l'*induzione a ritroso*: si parte dai nodi terminali, si risale verso la radice assegnando a ciascun nodo la scelta ottimale del giocatore che vi si muove, e si usa questa scelta come "payoff effettivo" del nodo padre.

Nell'esempio, il sottogame che inizia dopo che E ha scelto "Entra" ha un unico equilibrio: M sceglie A. Dunque l'unico equilibrio sottogame-perfetto è (Entra, A). L'algoritmo di Zermelo — scoperto nel contesto degli scacchi nel 1913 — è computazionalmente efficiente

(lineare nel numero di nodi) e produce un unico esito in giochi a informazione perfetta finiti.

La limitazione cruciale: l'equilibrio sottogame-perfetto è definito soltanto per i sottogame che iniziano da un nodo singolo. Nei giochi a informazione imperfetta, i sottogame sono rari o del tutto assenti, rendendo la raffinazione poco utile proprio nei contesti in cui le asimmetrie informative sono strutturalmente presenti — come la governance dell'AI.

L'equilibrio sequenziale di Kreps e Wilson

L'*equilibrio sequenziale* (Kreps e Wilson, 1982) generalizza la raffinazione ai giochi con insiemi informativi non singolari. Un equilibrio sequenziale è una coppia (σ, μ) dove σ sono strategie comportamentali e μ è un *sistema di credenze* — per ogni insieme informativo H , μ è una distribuzione di probabilità sui nodi di H che rappresenta le credenze del giocatore su dove si trova. La coppia (σ, μ) è un equilibrio sequenziale se soddisfa due condizioni.

La *consistenza* richiede che μ sia il limite di credenze bayesianamente coerenti per sequenze di perturbazioni che mettono probabilità positiva su tutti i nodi dell'albero. Questa condizione impedisce l'assegnazione arbitraria di credenze ai nodi fuori equilibrio — il punto dove la raffinazione morde davvero.

La *razionalità sequenziale* richiede che per ogni giocatore i e ogni insieme informativo H di i , la strategia σ_i massimizzi l'utilità attesa di i a partire da H , condizionatamente alle credenze μ in H e alle strategie degli altri giocatori.

La condizione di consistenza è il contributo tecnico chiave: esclude le credenze "artificiali" che supportano equilibri irrazionali ai nodi fuori equilibrio. La razionalità sequenziale garantisce che le strategie siano ottimali non soltanto globalmente ma ad ogni punto di decisione, condizionatamente alle credenze. Insieme, le due condizioni producono una nozione di equilibrio che è sia ragionevole dal punto di vista epistemico sia computazionalmente trattabile.

L'equilibrio trembling-hand perfetto di Selten

L'*equilibrio trembling-hand perfetto* (Selten, 1975) considera perturbazioni del gioco in cui ogni giocatore ha una probabilità $\epsilon > 0$ di scegliere ogni azione — gli "errori" dei giocatori. Un equilibrio è trembling-hand perfetto se è il limite di un equilibrio ϵ -perturbato quando $\epsilon \rightarrow 0$: soltanto le strategie che rimangono best response anche quando gli altri giocatori sbagliano con piccola probabilità sono ammissibili.

La gerarchia delle raffinazioni — Nash $\supseteq$ equilibrio sequenziale $\supseteq$ equilibrio sottogame-perfetto $\supseteq$ trembling-hand perfetto — riflette requisiti sempre più stringenti di razionalità ai nodi fuori equilibrio. Ogni raffinazione elimina alcuni equilibri di Nash ma può ancora lasciarne molteplici. La ricerca di raffinazioni più selettive ha prodotto una letteratura ampia e talvolta tecnica, ma il messaggio essenziale rimane: i meccanismi di governance che si reggono su minacce non credibili sono fragili, e la robustezza richiede che le strategie degli attori istituzionali siano eseguibili ex post, non solo annunciabili ex ante.

I.8 — Il prezzo dell'anarchia

Definizione e motivazione

Il *prezzo dell'anarchia* (Price of Anarchy, PoA), introdotto da Koutsoupias e Papadimitriou (1999) e sviluppato sistematicamente da Roughgarden e Tardos (2002), misura quantitativamente il costo del comportamento strategico non coordinato rispetto all'ottimo sociale:

$$\text{PoA}(G) = W(a^{**}) / \min_{a \in \text{NE}(G)} W(a)$$

dove $W(\cdot)$ è la misura di benessere sociale, a^{**} è il profilo di azioni socialmente ottimale, e $\text{NE}(G)$ è l'insieme degli equilibri di Nash del gioco G . Quando $\text{PoA} = 1$, ogni equilibrio è socialmente efficiente. Quando $\text{PoA} > 1$, vi è un costo strutturale della competizione non coordinata che cresce con PoA.

Il *prezzo dell'anarchia ottimistico* (Price of Stability, PoS) usa il massimo del benessere sugli equilibri al denominatore: misura il costo nel migliore equilibrio invece che nel peggiore. In molti giochi, $\text{PoS} = 1$ anche quando $\text{PoA} \gg 1$: esiste un equilibrio efficiente, ma i processi dinamici naturali potrebbero non convergere ad esso.

Il risultato di Roughgarden-Tardos

Il risultato fondamentale di Roughgarden e Tardos (2002) studia il routing del traffico nelle reti: ciascun utente sceglie il percorso che minimizza la propria latenza, ignorando l'effetto della propria scelta sulla latenza degli altri. Per giochi di congestione con funzioni di latenza polinomiali di grado al più d , il prezzo dell'anarchia soddisfa:

$$\text{PoA} \leq \Phi(d) = d / (d+1)^{\{(d+1)/d\}} \cdot (d+1)$$

dove $\Phi(d)$ dipende soltanto da d , non dalla topologia della rete né dalla distribuzione della domanda. Per funzioni di latenza lineari ($d=1$), $\text{PoA} \leq 4/3$.

Questo risultato è sorprendente per due ragioni. Il bound è *universale*: vale per qualsiasi rete, qualsiasi distribuzione della domanda, qualsiasi numero di utenti. Il bound è *stretto*: esistono reti in cui il PoA è esattamente $4/3$. Il "Paradosso di Braess" — in cui aggiungere un arco a una rete può peggiorare la latenza di tutti gli utenti — illustra come l'equilibrio di Nash del routing decentralizzato possa essere significativamente sub-ottimale.

La derivazione del bound per latenze lineari

Per reti con funzioni di latenza lineari $l_e(f_e) = a_e f_e + b_e$, il costo di un flusso f è $C(f) = \sum_e f_e l_e(f_e)$. In un equilibrio di Wardrop, ogni utente usa soltanto percorsi di latenza minima. Sia f^* il flusso ottimale e f^{eq} il flusso di equilibrio. Si ha:

$$C(f^{\text{eq}}) = \sum_e f_e^{\text{eq}} l_e(f_e^{\text{eq}})$$

Per la condizione di Wardrop, ogni utente non ha incentivo a deviare: il percorso che usa ha latenza minima. Questo implica:

$$\sum_e f_e^{\text{eq}} l_e(f_e^{\text{eq}}) \leq \sum_e f_e^* l_e(f_e^{\text{eq}}) \leq \sum_e f_e^* (a_e f_e^{\text{eq}} + b_e)$$

Per la linearità di l_e , $l_e(f_e^{\text{eq}}) = a_e f_e^{\text{eq}} + b_e \leq 2a_e f_e^{\text{eq}} + b_e / \dots$ e procedendo con la disuguaglianza di Young $ab \leq a^2/2 + b^2/2$, si ottiene il bound $\text{PoA} \leq 4/3$ per il caso lineare. $\square$

Applicazioni alla governance dell'AI

Per la governance dell'AI, il prezzo dell'anarchia è la misura naturale del costo sociale di un ecosistema non coordinato. I risultati di Roughgarden suggeriscono un'intuizione importante: ridurre il PoA non richiede eliminare la competizione ma modificare le funzioni di "latenza" — ovvero i costi marginali che ciascun attore percepisce per le proprie scelte di sviluppo. Meccanismi come la responsabilità legale per i danni da AI, le tasse pigouviane sui rischi sistemici, i premi per le pratiche di sicurezza, i costi di reputazione per comportamenti negligenti: tutti operano modificando le funzioni di costo percepite dagli agenti, avvicinando l'equilibrio non cooperativo all'ottimo sociale senza richiedere controllo centralizzato.

I.9 — Informazione asimmetrica: tipi, credenze e segnalazione

Il framework di Harsanyi

Il quadro sviluppato nelle sezioni precedenti assume che tutti i giocatori abbiano la stessa informazione sulla struttura del gioco. Questa assunzione è sistematicamente violata nel contesto della governance dell'AI: gli sviluppatori possiedono informazioni tecniche sui propri sistemi — capacità effettive, vulnerabilità, pratiche di sicurezza interne, valutazione dei rischi — che i regolatori, i deployer e gli utenti finali non possono accedere direttamente.

Harsanyi (1967-68) propose di modellare l'informazione asimmetrica associando a ciascun giocatore un *tipo privato* $\theta_i \in \Theta_i$, con la distribuzione congiunta $p(\theta) = p(\theta_1, \dots, \theta_n)$ *common knowledge* tra tutti i

giocatori — anche se la realizzazione θ_i è nota soltanto al giocatore i . Questa distinzione tra la distribuzione (pubblica) e la realizzazione (privata) è il cuore del framework bayesiano.

Un *equilibrio bayesiano di Nash* è un profilo di strategie tipo-dipendenti $\sigma^*(\theta) = (\sigma_1^*(\theta_1), \dots, \sigma_n^*(\theta_n))$ tale che ogni tipo θ_i di ogni giocatore i massimizza la propria utilità attesa condizionata alla propria informazione privata e alle strategie degli avversari:

$$\sigma_i^*(\theta_i) \in \operatorname{argmax}_{\{\sigma_i\}} E_{\{\theta_{-i}\} | \theta_i} [u_i(\sigma_i, \sigma_{-i}^*(\theta_{-i}), \theta_i, \theta_{-i})]$$

La dipendenza delle strategie dai tipi introduce una complessità che va oltre l'equilibrio di Nash standard: non si cerca soltanto un punto fisso nello spazio delle strategie, ma un punto fisso nello spazio delle *funzioni* dalle realizzazioni dei tipi alle azioni. Questo spazio è enormemente più grande e la dimostrazione dell'esistenza richiede tecniche più sofisticate.

Selezione avversa

La *selezione avversa*, analizzata da Akerlof nel celebre "The Market for Lemons" (1970, Nobel 2001), emerge quando una parte della transazione ha informazione superiore sulla qualità del bene scambiato. Il meccanismo è un circolo vizioso: i compratori, non potendo osservare la qualità, offrono un prezzo che riflette la qualità media del mercato; i venditori di beni superiori alla media ritirano le proprie offerte; la qualità media scende; il prezzo scende; ulteriori venditori di alta qualità escono; e così fino al collasso del mercato.

Nel contesto della governance dell'AI, la selezione avversa manifesta una dinamica analoga nella certificazione della sicurezza: se la certificazione non è sufficientemente discriminante, le organizzazioni con sistemi

effettivamente pericolosi possono soddisfare i requisiti attraverso comportamenti di test-gaming, abbassando il valore informativo della certificazione e potenzialmente spingendo le organizzazioni con sistemi genuinamente sicuri a non investire nell'ottenimento della certificazione — perché il segnale non è credibile.

Segnalazione e single crossing

La *segnalazione*, analizzata da Spence (1973, Nobel 2001), emerge come meccanismo con cui i tipi di alta qualità comunicano credibilmente la propria qualità attraverso azioni costose che i tipi di bassa qualità non hanno incentivo a imitare. La condizione strutturale che rende possibile la segnalazione è la *single crossing*: il costo marginale del segnale è inferiore per i tipi di alta qualità.

Se questa condizione è soddisfatta, esiste un livello di segnale e^* sufficientemente alto da soddisfare entrambi i vincoli di compatibilità degli incentivi:

$$\begin{aligned} v_H \cdot B(e^*) - C_H(e^*) &\geq v_H \cdot B(0) \text{ [il tipo alto preferisce segnalare]} \\ v_L \cdot B(e^*) - C_L(e^*) &< v_L \cdot B(0) \text{ [il tipo basso non trova conveniente imitare]} \end{aligned}$$

In un equilibrio *separante*, il tipo alto sceglie e^* e il tipo basso sceglie il segnale minimo, e il mercato li tratta diversamente in base al segnale osservato. In un equilibrio *pooling*, entrambi i tipi inviano lo stesso segnale e il mercato li tratta in modo identico — producendo meno informazione e in genere meno efficienza allocativa.

Per la governance dell'AI, la struttura degli equilibri separanti suggerisce che i meccanismi dovrebbero essere progettati per facilitare la separazione — abbassando i costi del segnale per il tipo alto (chi ha

effettivamente buone pratiche di sicurezza) senza abbassarli per il tipo basso. Gli standard di processo soddisfano questa condizione meglio degli standard di risultato puramente basati su benchmark statici: le organizzazioni con buone pratiche le soddisfano a costo relativamente basso, mentre quelle che non le hanno devono costruirle da zero — ed è questa differenza di costo a produrre la separazione informativa.

La condizione di single crossing ha un'implicazione strutturale che merita enfasi. Essa non è un'assunzione arbitraria introdotta per ottenere risultati gradevoli: è una condizione sulla geometria delle curve di costo dei diversi tipi che può essere verificata empiricamente (in linea di principio) e che riflette una differenza qualitativa tra i tipi — non una differenza di quantità ma di struttura dei costi marginali. Quando la single crossing è violata, i tipi non sono separabili attraverso la segnalazione, e il regolatore deve ricorrere ad altri meccanismi di rivelazione dell'informazione.

IL MECCANISMO DESIGN

II.1 — L'inversione del problema: dalla teoria positiva alla teoria normativa delle istituzioni

La rivoluzione concettuale

La teoria dei giochi classica affronta l'interazione strategica nella sua direzione naturale: *dato* un gioco — dati i giocatori, le loro azioni disponibili, le loro funzioni di utilità, la struttura informativa — si analizzano gli equilibri, si valutano le proprietà di efficienza e stabilità, si descrivono le previsioni comportamentali. È una teoria *positiva*: descrive

ciò che accade, o meglio ciò che accadrebbe se gli agenti fossero razionali nel senso specificato dalla teoria.

Il meccanismo design — chiamato da Leonid Hurwicz "the theory of incentives" e talvolta "theory of games in reverse" — inverte sistematicamente questa direzione. *Dato* l'esito desiderato, si progettano le regole del gioco affinché quell'esito emerga come equilibrio del comportamento razionale degli agenti. È una teoria *normativa* e *ingegneristica*: prescrive come costruire le istituzioni che producono certi esiti.

Questa inversione non è una semplice tecnica analitica: è una rivoluzione concettuale di prima grandezza. Il meccanismo design tratta le istituzioni — i sistemi di regole, di incentivi, di procedure decisionali — come oggetti *progettabili*, non come dati esogeni su cui la teoria può soltanto riflettere post-hoc. Le aste, i mercati, le votazioni, i contratti, le procedure di allocazione delle risorse pubbliche, le regole di governance delle organizzazioni: tutto questo è, nella prospettiva del meccanismo design, *architettura di incentivi*, costruita deliberatamente per produrre certi esiti a partire dal comportamento strategico di agenti con obiettivi privati e informazione asimmetrica.

Vale la pena apprezzare la portata di questa svolta. Prima del meccanismo design, l'analisi economica delle istituzioni era prevalentemente un'analisi dei loro effetti: dato che esiste questa regola, qual è l'equilibrio che produce? Il meccanismo design pone la domanda opposta: quale regola produce l'equilibrio desiderato? L'apparente simmetria nasconde una difficoltà di un ordine di grandezza superiore. Invertire la direzione dell'analisi significa lavorare in uno spazio enorme — lo spazio di tutti i possibili meccanismi — e trovare quelli che soddisfano simultaneamente più criteri: efficienza allocativa,

compatibilità degli incentivi alla rivelazione, partecipazione volontaria, sostenibilità finanziaria. Il risultato fondamentale del campo — il principio di rivelazione, discusso nella sezione seguente — è precisamente la tecnica che rende questo spazio di ricerca trattabile.

Il riconoscimento del Nobel del 2007, assegnato congiuntamente a Hurwicz, Maskin e Myerson, ha sancito la maturità e l'importanza pratica di questo programma. Le applicazioni si dispiegano su un'ampiezza sorprendente: dall'allocazione delle frequenze radio attraverso aste multi-round (Milgrom e Wilson, Nobel 2020) al mercato dei reni senza prezzi (Roth e Shapley, Nobel 2012), dalle aste per i titoli del tesoro all'assegnazione degli studenti alle scuole nelle grandi città metropolitane. In ciascun caso, l'essenza del problema è la stessa: come progettare le regole di interazione affinché agenti con informazione privata rivelino volontariamente quella informazione e si coordinino verso esiti socialmente desiderabili, senza la necessità di sorveglianza totale, coercizione diretta o allineamento dei valori.

Il framework formale

Un *meccanismo* è una coppia (M, g) dove $M = \times_i M_i$ è lo spazio dei messaggi — ogni giocatore i invia un messaggio $m_i \in M_i$ — e $g : M \rightarrow X$ è la funzione di risultato che mappa i messaggi nell'insieme degli esiti X . Un meccanismo induce un *gioco di implementazione*: i giocatori scelgono strategicamente i propri messaggi, e l'esito è determinato da g applicata ai messaggi scelti.

L'ambiente è descritto da un profilo di tipi privati $\theta = (\theta_1, \dots, \theta_n) \in \times_i \Theta_i$ e da un insieme di esiti X . Nell'ambiente *quasi-lineare standard* — il più studiato e il più ricco di risultati — $X = \{x \in \mathbb{R}^+ : \text{allocazione fisica}\} \times \mathbb{R}^n$ (trasferimenti monetari), e l'utilità del giocatore i di tipo θ_i è:

$$u_i(x, t_i; \theta_i) = v_i(x, \theta_i) + t_i$$

dove v_i è la valorizzazione dell'allocazione fisica e t_i è il trasferimento monetario ricevuto (positivo se ricevuto, negativo se pagato). La linearità in t_i — l'assunzione quasi-lineare — è ciò che rende possibile la "monetizzazione" degli incentivi: i trasferimenti monetari possono compensare esattamente le perdite di utilità derivanti da allocazioni sub-ottimali per tipi specifici.

Una *funzione di scelta sociale* (SCC) $f : \times_i \Theta_i \rightarrow X$ specifica l'esito desiderato per ogni profilo di tipi. Un meccanismo (M, g) *implementa* f se esiste un equilibrio del gioco indotto in cui l'esito prodotto coincide con $f(\theta)$ per ogni θ . Tre diversi concetti di implementazione si distinguono per il grado di robustezza richiesto.

L'implementazione in strategie dominanti (DSIC) richiede che per ogni giocatore i , per ogni tipo θ_i , e per ogni comportamento degli altri giocatori, il giocatore i massimizzi la propria utilità inviando il messaggio ottimale indipendentemente da cosa fanno gli altri. È la forma più robusta: non richiede credenze sui tipi altrui, non dipende da assunzioni sulla razionalità degli altri, non richiede coordinazione sull'equilibrio. È la forma di implementazione che funziona anche se i giocatori non si fidano l'uno dell'altro, non condividono alcun prior comune, e non hanno capacità computazionali illimitate.

L'implementazione in equilibrio bayesiano di Nash (BIC) richiede che esista un equilibrio bayesiano di Nash del gioco indotto che produce l'esito $f(\theta)$. Richiede che gli agenti condividano un prior comune e si coordinino sull'equilibrio specifico. Permette meccanismi più efficienti del DSIC perché rilassa i requisiti di robustezza: il progettista sfrutta la struttura del prior per estrarre più informazione privata.

L'implementazione **ex-post** richiede che l'esito sia $f(\theta)$ per ogni realizzazione dei tipi, senza rimpianti *ex-post*. Bergemann e Morris (2005) hanno mostrato che, sotto certe condizioni, l'implementazione *ex-post* è equivalente all'implementazione in strategie dominanti — un risultato che rafforza considerevolmente la motivazione per la classe dei meccanismi DSIC.

II.2 — Il principio di rivelazione

Enunciato e dimostrazione

Il teorema più importante nel fondamento del meccanismo design è il *principio di rivelazione*, che riduce lo spazio dei meccanismi da considerare — potenzialmente illimitato — a una classe strutturata e analiticamente trattabile.

Teorema (Principio di Rivelazione): Se una SCC f è implementabile da un meccanismo arbitrario (M, g) con un equilibrio σ^* , allora f è implementabile da un meccanismo diretto $(\Theta, \tilde{f})$ in cui lo spazio dei messaggi coincide con lo spazio dei tipi e la rivelazione veritiera è un equilibrio.

Dimostrazione. Dato il meccanismo (M, g) con equilibrio $\sigma^* = (\sigma_{-1}^*, \dots, \sigma_n^*)$, costruiamo il meccanismo diretto rivelante $(\Theta, \tilde{f})$ come segue:

$$\tilde{f}(\theta) = g(\sigma_{-1}^*(\theta_{-1}), \dots, \sigma_n^*(\theta_n))$$

Il meccanismo esegue implicitamente le strategie di equilibrio del meccanismo originale: se il giocatore i dichiara il tipo θ_i , il meccanismo

si comporta come se il giocatore stesse inviando il messaggio $\sigma_i^*(\theta_i)$ nel meccanismo originale.

La rivelazione veritiera è un equilibrio perché: se il giocatore i di tipo θ_i potesse migliorare dichiarando $\theta_i \neq \theta_i$ nel meccanismo diretto, questo implicherebbe che avrebbe potuto migliorare inviando il messaggio $\sigma_i^*(\theta_i)$ invece di $\sigma_i^*(\theta_i)$ nel meccanismo originale — contraddicendo l'ipotesi che σ^* sia un equilibrio di quel meccanismo. $\square$

Implicazioni e limiti

Il principio ha una conseguenza di portata straordinaria: la ricerca del meccanismo ottimale può essere ristretta ai meccanismi diretti incentive-compatibili, caratterizzati attraverso condizioni algebriche esplicite sulle funzioni di allocazione e pagamento. Invece di ottimizzare su uno spazio infinito-dimensionale di meccanismi arbitrari — aste con strutture diverse, contratti con clausole varie, procedure di approvazione graduali — il problema si riduce a ottimizzare una funzione obiettivo soggetta a sistemi di disuguaglianze, un problema molto più strutturato e spesso risolvibile analiticamente.

Un primo caveat importante: il principio vale come "non perdita di generalità" nella caratterizzazione teorica degli esiti implementabili, ma non implica che i meccanismi diretti siano preferibili in pratica. I meccanismi indiretti — aste dinamiche con offerte iterate, negoziazioni multi-round, processi di approvazione graduali — hanno spesso proprietà pratiche superiori: sono più semplici da comprendere per i partecipanti, consentono di aggiornare le credenze nel corso del processo, e producono esiti vicini all'ottimale anche in presenza di razionalità limitata degli agenti. Il principio di rivelazione dice soltanto

che, ai fini della caratterizzazione teorica degli esiti implementabili, non occorre guardare oltre i meccanismi diretti veritieri.

Un secondo caveat, meno frequentemente discusso, riguarda i *meccanismi non mediati*. Il principio assume che esista un mediatore affidabile che riceve le dichiarazioni dei tipi e implementa l'esito. In contesti in cui tale mediatore non esiste o non è credibile — come la governance internazionale in assenza di autorità sovranazionale — il principio non si applica direttamente: i giocatori devono raggiungere coordinazione senza mediazione centrale, e questo riduce lo spazio degli esiti implementabili. La letteratura sui meccanismi senza mediatore — *unmediated mechanisms* — mostra che la perdita può essere sostanziale in certi giochi.

Un terzo caveat, di natura più fondamentale, riguarda la *complessità computazionale della rivelazione*. In contesti con spazi di tipi molto ricchi — come quelli della governance AI, dove il "tipo" di un sistema AI potrebbe essere un oggetto di elevata dimensionalità — dichiarare il tipo veritiero può richiedere una capacità computazionale e comunicativa proibitiva. Il principio di rivelazione è un risultato di esistenza, non di computabilità.

II.3 — Il meccanismo VCG: derivazione e proprietà

Il problema dell'efficienza con incentivi

Il problema fondamentale che il meccanismo design cerca di risolvere nella sua versione più elementare è: come produrre l'esito socialmente

efficiente quando le valorizzazioni degli agenti sono informazione privata? Si vuole massimizzare il benessere sociale:

$$W(x, \theta) = \sum_i v_i(x, \theta_i)$$

con la complicazione che i tipi privati θ_i non sono osservabili dal progettista. Se il progettista conoscesse i tipi, potrebbe semplicemente calcolare l'esito efficiente. Non li conosce, e gli agenti hanno incentivo a mentire se questo migliora la propria allocazione.

La tensione sembra insolubile: l'efficienza richiede di considerare le valorizzazioni reali di tutti, ma le valorizzazioni reali sono private. Il meccanismo VCG risolve questa tensione con un'idea elegante: progettare i trasferimenti monetari in modo che per ogni agente, dichiarare il vero sia l'azione che massimizza la propria utilità *inclusi* i trasferimenti, indipendentemente da cosa fanno gli altri.

Il meccanismo di Vickrey–Clarke–Groves

Il meccanismo di Vickrey (1961) — Clarke (1971) — Groves (1973) funziona come segue. Ciascun agente i dichiara le proprie valorizzazioni $\hat{v}_i(\cdot)$. Il meccanismo:

(a) Sceglie l'esito che massimizza il benessere dichiarato totale:

$$x^*(\hat{v}) = \operatorname{argmax}_{x \in X} \sum_i \hat{v}_i(x)$$

(b) Assegna a ogni agente i un trasferimento monetario:

$$t_i(\hat{v}) = \sum_{j \neq i} \hat{v}_j(x^*(\hat{v})) - h_i(\hat{v}_{-i})$$

dove h_i è una funzione arbitraria delle dichiarazioni degli *altri* agenti — non di i stesso.

Teorema: La dichiarazione veritiera $\hat{v}_i = v_i$ è una strategia debolmente dominante per ogni agente i .

Dimostrazione. L'utilità netta dell'agente i di tipo v_i che dichiara $\hat{v}_i$, quando gli altri dichiarano $\hat{v}_{-i}$, è:

$$\begin{aligned}\Pi_i &= v_i(x^*(\hat{v}_i, \hat{v}_{-i})) + t_i(\hat{v}_i, \hat{v}_{-i}) \\ &= v_i(x^*(\hat{v}_i, \hat{v}_{-i})) + \sum_{j \neq i} \hat{v}_j(x^*(\hat{v}_i, \hat{v}_{-i})) - h_i(\hat{v}_{-i})\end{aligned}$$

Il termine $h_i(\hat{v}_{-i})$ non dipende da $\hat{v}_i$. Massimizzare Π_i rispetto a $\hat{v}_i$ è dunque equivalente a massimizzare:

$$v_i(x^*(\hat{v}_i, \hat{v}_{-i})) + \sum_{j \neq i} \hat{v}_j(x^*(\hat{v}_i, \hat{v}_{-i}))$$

Se l'agente dichiara veritieramente $\hat{v}_i = v_i$, la scelta x^* massimizza $\sum_j \hat{v}_j(x)$ che, con $\hat{v}_i = v_i$, coincide con $\sum_j v_j(x)$ — il benessere sociale reale. Nessuna dichiarazione alternativa $\hat{v}_i \neq v_i$ può produrre un esito con benessere sociale reale maggiore (per definizione di x^*), e quindi nessuna dichiarazione alternativa può aumentare l'utilità netta di i . $\square$

La profondità di questo risultato risiede nel fatto che il VCG trasforma il problema dell'agente da "massimizzare la propria quota" a "massimizzare il benessere collettivo": dichiarando il vero, l'agente sceglie implicitamente l'allocatione che massimizza il totale, e il meccanismo lo ricompensa esattamente di quanto il totale vale. L'agente diventa, in un senso preciso, l'ottimizzatore del benessere sociale.

Il meccanismo pivotal di Clarke

La scelta più comune di h_i — che porta al meccanismo *pivotal* di Clarke — è:

$$h_i(\hat{v}_{-i}) = \max_{x \in X} \sum_{j \neq i} \hat{v}_j(x)$$

Con questa scelta, il trasferimento t_i è la differenza tra il benessere degli altri agenti con l'agente i (che contribuisce la propria valorizzazione alla scelta di x^*) e il benessere ottimale degli altri agenti *in assenza di i* . L'interpretazione economica è cristallina: l'agente i paga l'*esternalità negativa* che impone agli altri — la riduzione del benessere altrui causata dalla necessità di tenere conto delle sue preferenze nella scelta collettiva. Se i è *pivotal* — se la sua presenza cambia l'esito ottimale rispetto a quello che si sceglierebbe senza di lui — paga una tassa positiva pari all'esternalità che impone. Se non è pivotal, non paga nulla.

Esempio: Asta VCG per un oggetto singolo con tre compratori aventi valorizzazioni $v_1 = 100$, $v_2 = 70$, $v_3 = 50$. L'oggetto va al compratore 1 (che ha la valorizzazione più alta). Il suo pagamento è:

$$t_1 = \max(70, 50) - \max(0) = 70$$

dove $\max(70, 50)$ è il benessere ottimale degli altri con l'oggetto assegnato al migliore tra loro (il compratore 2, con valorizzazione 70), e $\max(0)$ è il benessere ottimale degli altri se l'oggetto non venisse allocato a nessuno dei presenti (zero, poiché senza il compratore 1 il migliore è il compratore 2 che ottiene l'oggetto gratuitamente). Il meccanismo VCG coincide con l'asta di secondo prezzo di Vickrey: il vincitore paga il prezzo del secondo miglior offerente.

Proprietà e limiti strutturali

Il VCG soddisfa simultaneamente tre proprietà che sembrano in tensione: produce l'esito socialmente efficiente, rende la verità strategia debolmente dominante, e non richiede che il progettista conosca i tipi

privati degli agenti. Questo lo rende il riferimento teorico per qualsiasi problema di allocazione efficiente con informazione privata.

Tuttavia il VCG presenta limitazioni strutturali importanti che è necessario esaminare con attenzione.

Non budget-bilanciamento: In generale, $\sum_i t_i \leq 0$ — il meccanismo brucia denaro o richiede sussidi esterni. Questo è un limite reale in molti contesti di governance: un meccanismo che richiede trasferimenti netti dal pubblico erario non è auto-sostenibile senza una fonte esterna di finanziamento. L'intuizione economica è che le rendite informative pagate agli agenti per indurre la rivelazione veritiera eccedono le tasse che essi pagano per le esternalità — il costo dell'informazione privata non può essere recuperato internamente.

Rendite informative: I tipi con alta valorizzazione ricevono rendite informative — payoff superiori al loro costo di opportunità — che sono il prezzo necessario per indurre la rivelazione veritiera. Questi trasferimenti redistribuiscono il surplus dagli acquirenti verso i tipi ad alta valorizzazione in modo che può risultare politicamente inaccettabile in certi contesti di governance.

Manipolazione collettiva: Sebbene il VCG sia robusto alle deviazioni unilaterali, non è robusto alle deviazioni collettive: un gruppo di agenti coordinati può talvolta manipolare il meccanismo a proprio vantaggio attraverso strategie congiunte non previste dall'analisi unilaterale. Ausubel e Milgrom (2006) analizzano questo problema nei dettagli.

Complessità computazionale: In contesti con molti agenti e molti esiti possibili, calcolare $x^*(\hat{v}) = \operatorname{argmax} \sum_i \hat{v}_i(x)$ può essere computazionalmente intrattabile. Le aste combinatoriali — dove i compratori hanno valorizzazioni per pacchetti di oggetti — sono

NP-difficili in generale, e il problema si ripropone in ogni contesto di governance con strutture di allocazione complesse.

II.4 — Il teorema di Green-Laffont: unicità della classe VCG

Il VCG non è soltanto un meccanismo efficiente-DSIC particolarmente elegante: è l'*unica* famiglia di meccanismi con queste proprietà. Questo risultato è importante perché chiude lo spazio della ricerca: chiunque voglia efficienza con massima robustezza non ha alternative.

Teorema (Green e Laffont, 1977): In un ambiente quasi-lineare con spazio dei tipi sufficientemente ricco — nel senso che per ogni agente i e ogni coppia di esiti $x, x' \in X$ esistono tipi θ_i e θ'_i tali che $v_i(x, \theta_i) > v_i(x', \theta_i)$ e $v_i(x', \theta'_i) > v_i(x, \theta'_i)$ — qualsiasi meccanismo che implementa la funzione di scelta efficiente in strategie dominanti deve avere trasferimenti della forma VCG:

$$t_i(\hat{v}) = \sum_{j \neq i} \hat{v}_j(x^*(\hat{v})) - h_i(\hat{v}_{-i})$$

per qualche funzione h_i arbitraria.

Schema della dimostrazione. Consideriamo due tipi θ_i e θ'_i distinti per il giocatore i . La condizione DSIC richiede che per ogni tipo e ogni comportamento degli altri, il tipo θ_i preferisca dichiarare il vero:

$$v_i(x^*(\theta_i, v_{-i})) + t_i(\theta_i, v_{-i}) \geq v_i(x^*(\theta'_i, v_{-i})) + t_i(\theta'_i, v_{-i})$$

e analogamente il tipo θ'_i preferirebbe dichiarare il proprio tipo:

$$v_i(x^*(\theta_i, v_{-i})) + t_i(\theta_i, v_{-i}) \geq v_i(x^*(\theta_i, v_{-i})) + t_i(\theta_i, v_{-i})$$

Sommando le due disuguaglianze si ottiene un'uguaglianza, che implica che entrambe siano equazioni. Applicando questa condizione a coppie di tipi scelte opportunamente — sfruttando la ricchezza dello spazio dei tipi per costruire valorizzazioni che rivelano la struttura dei trasferimenti — si mostra che la differenza $t_i(\theta_i, v_{-i}) - t_i(\theta_i, v_{-i})$ è determinata dalla differenza delle somme delle valorizzazioni degli altri in corrispondenza delle allocazioni ottimali rispettive. Iterando questo argomento si ottiene che i trasferimenti devono avere la struttura VCG. $\square$

Il teorema di Green-Laffont delimita con precisione lo spazio della "prima scelta" nel meccanismo design. Le limitazioni del VCG — il non budget-bilanciamento, le rendite informative, la vulnerabilità alla manipolazione collettiva — non sono difetti contingenti di un meccanismo specifico: sono caratteristiche *necessarie* di qualsiasi meccanismo efficiente in strategie dominanti. Questo è il messaggio centrale del risultato: non si può avere efficienza con massima robustezza e al contempo budget-bilanciamento e assenza di rendite. Le limitazioni del VCG non sono progettabili via — sono il prezzo strutturale della combinazione di proprietà che il VCG realizza.

II.5 — Il teorema di Myerson-Satterthwaite: il costo irrinunciabile dell'informazione privata

Il problema e l'impossibilità

Il teorema di Myerson e Satterthwaite (1983) costituisce il primo e più influente limite strutturale del meccanismo design, mostrando l'incompatibilità di quattro proprietà che sembrano tutte ragionevoli nel contesto della contrattazione bilaterale. Il suo messaggio è che l'informazione privata impone un costo irrinunciabile sull'efficienza degli scambi — un costo che nessun meccanismo, per quanto ingegnoso, può eliminare.

Il problema: un venditore con costo privato θ_S distribuito uniformemente su $[c, d]$ e un compratore con valorizzazione privata θ_B distribuito uniformemente su $[a, b]$ vogliono commerciare in modo efficiente. Le quattro proprietà desiderabili sono:

Efficienza ex-post (EFF): commerciare se e solo se $\theta_B \geq \theta_S$ — si realizza lo scambio esattamente quando è socialmente desiderabile.

Compatibilità bayesiana degli incentivi (BIC): per entrambi i giocatori, dichiarare il tipo vero è best response quando l'altro fa lo stesso.

Razionalità individuale interim (IIR): ogni tipo di ogni giocatore ottiene utilità attesa non negativa partecipando — nessuno è costretto a partecipare subendo una perdita attesa.

Budget-bilanciamento (BB): i trasferimenti monetari netti del meccanismo sono zero. Il meccanismo non richiede sussidi da una terza parte.

Teorema (Myerson-Satterthwaite, 1983): Nessun meccanismo può soddisfare simultaneamente EFF, BIC, IIR e BB quando i supporti dei tipi si sovrappongono ($a < d$ e $c < b$).

Dimostrazione per envelope theorem

La dimostrazione sfrutta l'*envelope theorem* della teoria della rivelazione. Per ogni meccanismo BIC, la derivata del payoff rispetto al tipo è determinata dalla regola di allocazione — un fatto che emerge dalla condizione di ottimo del problema di massimizzazione.

Sia $q(\theta_B, \theta_S) \in [0,1]$ la probabilità di commercio. Definiamo le probabilità marginali di commercio:

$$Q_B(\theta_B) = E_{\{\theta_S\}}[q(\theta_B, \theta_S)] \quad Q_S(\theta_S) = E_{\{\theta_B\}}[q(\theta_B, \theta_S)]$$

Per la BIC e l'envelope theorem, i payoff interim — il payoff atteso condizionato al proprio tipo — soddisfano:

$$V_B(\theta_B) = V_B(a) + \int_a^{\theta_B} Q_B(t) dt \quad [\text{compratore}] \quad V_S(\theta_S) = V_S(d) - \int_{\theta_S}^d Q_S(t) dt \quad [\text{venditore}]$$

La BIC implica anche che Q_B sia non-decrescente e Q_S sia non-crescente: i tipi più alti del compratore commerciano con probabilità maggiore, i tipi più bassi del venditore commerciano con probabilità maggiore.

Per budget-bilanciamento, i trasferimenti si annullano. Il surplus totale atteso distribuito ai partecipanti è $E[V_B(\theta_B)] + E[V_S(\theta_S)]$. Integrando per parti il termine del compratore con distribuzioni uniformi su $[a,b]$ e $[c,d]$:

$$E[V_B(\theta_B)] = V_B(a) + E[(1-F_B(\theta_B))/f_B(\theta_B) \cdot Q_B(\theta_B)] = V_B(a) + E[\varphi_B(\theta_B) \cdot q(\theta_B, \theta_S)]$$

dove $\varphi_B(\theta_B) = \theta_B - (1-F_B(\theta_B))/f_B(\theta_B)$ è il valore virtuale del compratore. Analogamente per il venditore, con valore virtuale $\varphi_S(\theta_S) = \theta_S + F_S(\theta_S)/f_S(\theta_S)$.

Per budget-bilanciamento e IIR:

$$E[(\theta_B - \theta_S) \cdot 1_{\{\theta_B \geq \theta_S\}}] = V_B(a) + V_S(d) + E[(\varphi_B(\theta_B) - \varphi_S(\theta_S)) \cdot 1_{\{\theta_B \geq \theta_S\}}]$$

Poiché $\varphi_B(\theta_B) < \theta_B$ e $\varphi_S(\theta_S) > \theta_S$, la differenza $\varphi_B - \varphi_S < \theta_B - \theta_S$, dunque:

$$E[(\varphi_B(\theta_B) - \varphi_S(\theta_S)) \cdot 1_{\{\theta_B \geq \theta_S\}}] < E[(\theta_B - \theta_S) \cdot 1_{\{\theta_B \geq \theta_S\}}]$$

Questo implica $V_B(a) + V_S(d) > 0$ — ma IIR richiede entrambe non negative, dunque il sistema è strettamente consistente fin qui. Il punto critico emerge quando i supporti si sovrappongono: la differenza tra i due lati è:

$$E[((1-F_B(\theta_B))/f_B(\theta_B) + F_S(\theta_S)/f_S(\theta_S)) \cdot 1_{\{\theta_B \geq \theta_S\}}] > 0$$

Questo termine è strettamente positivo, e l'uguaglianza richiesta da BB e IIR non può essere soddisfatta simultaneamente con EFF. $\square$

L'intuizione economica e le implicazioni per la governance

L'intuizione è la seguente. Per rivelare credibilmente il proprio tipo basso (costo basso), il venditore deve ricevere una rendita informativa che lo incentivi a non fingere un tipo ancora più basso (che porterebbe a un prezzo ancora più alto). Analogamente il compratore per rivelare un tipo alto. Queste rendite informative sono il costo della compatibilità degli

incentivi — il prezzo dell'informazione privata. Quando i supporti si sovrappongono, la massa di probabilità nella zona di "incertezza" è sufficientemente grande che le rendite necessarie superano il surplus atteso dello scambio efficiente in quella zona.

Il teorema di Myerson-Satterthwaite ha implicazioni dirette per la certificazione della sicurezza AI. Il problema ha esattamente la struttura della contrattazione bilaterale: lo sviluppatore conosce la sicurezza effettiva del proprio sistema (tipo privato), il regolatore vuole un'allocazione efficiente (certificare i sistemi sicuri, bloccare quelli pericolosi), e la partecipazione deve essere volontaria.

Il teorema dice che nessun meccanismo di certificazione può essere simultaneamente efficiente, incentive-compatible, individualmente razionale e auto-finanziante. La scelta di quale proprietà rilassare è una decisione di *policy* con conseguenze concrete: rilassare l'efficienza significa accettare errori sistematici nella classificazione dei sistemi; rilassare la razionalità individuale significa rendere obbligatoria la certificazione; rilassare il budget-bilanciamento significa finanziare l'infrastruttura di governance come bene pubblico; rilassare la compatibilità degli incentivi significa accettare che gli sviluppatori abbiano incentivi a manipolare il processo. Nessuna di queste opzioni è gratuita, e il teorema dice che non si può evitare di scegliere.

II.6 — L'asta ottimale di Myerson: massimizzare il ricavo

Il valore virtuale

Il contributo fondamentale di Myerson del 1981 risponde a una domanda diversa da quella dell'efficienza: quale meccanismo massimizza l'obiettivo del *progettista* — il ricavo del venditore — soggetto ai soli vincoli di compatibilità degli incentivi e partecipazione volontaria, rinunciando esplicitamente al budget-bilanciamento?

Il passo cruciale è la riformulazione attraverso il *valore virtuale*. Dall'envelope theorem applicato alla BIC, il ricavo atteso del venditore da ogni compratore i è:

$$E[t_i(\theta)] = E[v_i(x^*(\theta), \theta_i)] - \int Q_i(t_i) dt_i - V_i(\theta_i)$$

Integrando per parti si ottiene:

$$\text{Ricavo} = E[\varphi_i(\theta_i) \cdot q_i(\theta)]$$

dove il *valore virtuale* del compratore i di tipo θ_i è:

$$\varphi_i(\theta_i) = \theta_i - (1 - F_i(\theta_i)) / f_i(\theta_i)$$

Il valore virtuale "corregge" la valorizzazione reale θ_i sottraendo il termine informativo $(1-F_i)/f_i$ — la rendita informativa che il compratore di tipo θ_i deve ricevere affinché i tipi più alti abbiano incentivo a rivelare il vero. Il valore virtuale è inferiore al valore reale: il monopolista non può estrarre l'intera valorizzazione perché deve lasciare rendite informative ai tipi alti per indurli alla rivelazione.

La regola di allocazione ottimale e il prezzo di riserva

Massimizzare il ricavo atteso equivale a massimizzare $E[\sum_i \varphi_i(\theta_i) \cdot q_i(\theta)]$ soggetto ai vincoli di fattibilità. Sotto la condizione di *regolarità* — $\varphi_i(\theta_i)$ è crescente in θ_i , condizione soddisfatta per le distribuzioni uniformi e molte distribuzioni log-concave — la regola ottimale è:

$q_i^*(\theta) = 1$ se $\varphi_i(\theta_i) = \max_j \varphi_j(\theta_j)$ e $\varphi_i(\theta_i) \geq 0$ $q_i^*(\theta) = 0$ altrimenti

La regola assegna l'oggetto al compratore con il valore virtuale massimo tra quelli non negativi, e non alloca a nessuno se tutti i valori virtuali sono negativi. Questo incorpora un *prezzo di riserva ottimale* $r_i = \varphi_i^{-1}(0)$: il venditore non vende mai al di sotto di questo prezzo, anche se tutti i compratori hanno valorizzazioni positive. Per la distribuzione uniforme su $[0,1]$, $\varphi(\theta) = 2\theta - 1$, dunque $r = 1/2$: il venditore fissa il prezzo minimo a metà del valore massimo possibile, anche se questo implica non vendere con probabilità positiva.

Il teorema di equivalenza dei ricavi

Il teorema di equivalenza dei ricavi (Myerson 1981; Riley e Samuelson 1981) è uno dei risultati più eleganti di tutta la teoria economica: stabilisce che, in condizioni sufficientemente generali, formati d'asta radicalmente diversi producono lo stesso ricavo atteso al venditore.

Teorema: Consideriamo qualsiasi due meccanismi BIC con n compratori con valorizzazioni i.i.d. da F tali che: (1) assegnino l'oggetto allo stesso compratore per ogni profilo θ ; (2) diano lo stesso payoff al tipo con valorizzazione minima. Allora generano lo stesso ricavo atteso.

Dimostrazione. Per BIC e l'envelope theorem, il payoff del compratore i di tipo θ_i è:

$$V_i(\theta_i) = V_i(\underline{\theta}) + \int_{\underline{\theta}}^{\theta_i} Q_i(t) dt$$

Le condizioni (1) e (2) implicano che Q_i e $V_i(\underline{\theta})$ sono identici nei due meccanismi. Dunque $V_i(\theta_i)$ è identico. Il pagamento atteso $E[t_i(\theta_i)]$

$= E[\theta_i \cdot q_i(\theta)] - V_i(\theta_i)$ è identico. Sommando su tutti i compratori, i ricavi coincidono. $\square$

L'asta inglese a prezzi crescenti, l'asta olandese a prezzi decrescenti, l'asta di primo prezzo in busta chiusa e l'asta di secondo prezzo di Vickrey: in condizioni simmetriche con compratori indipendenti e identicamente distribuiti, tutte producono lo stesso ricavo atteso in equilibrio. L'implicazione per la governance è profonda: il formato di un meccanismo non conta in sé — ciò che conta è la struttura degli incentivi che crea. Due meccanismi di governance che producono la stessa funzione di allocazione degli incentivi genereranno gli stessi esiti comportamentali in equilibrio, anche se appaiono radicalmente diversi in superficie.

II.7 — Meccanismo design multi-dimensionale e i suoi limiti

La maledizione delle dimensioni

Il meccanismo design diventa drasticamente più complesso quando i tipi privati degli agenti sono multi-dimensionali. Questa è la norma nei problemi reali: uno sviluppatore di AI ha informazione privata sia sulle capacità del proprio sistema, sia sui costi di sviluppo, sia sui rischi sistemici, sia sulle intenzioni di deployment — un vettore di tipo, non uno scalare.

Nel caso uni-dimensionale, la BIC implica che la funzione di allocazione attesa $Q_i(\theta_i)$ sia *monotona non-decrescente* in θ_i — una condizione

semplice e verificabile. Nel caso multi-dimensionale, con tipo $\theta_i \in \mathbb{R}^k$, la BIC implica che il gradiente dell'utilità del tipo rispetto al tipo stesso soddisfi una condizione di simmetria e le condizioni di integrabilità di un potenziale:

$$\partial Q_i^j / \partial \theta_i^l = \partial Q_i^l / \partial \theta_i^j \text{ per ogni } j, l$$

dove Q_i^j è la derivata dell'utilità attesa rispetto alla j -esima componente del tipo. Questa condizione di *irrotazionalità* nel senso della teoria dei flussi è molto più complessa da verificare e garantire, e l'ottimo in genere non ha una forma analitica semplice.

Il principale risultato negativo, dovuto a Rochet e Choné (1998) e approfondito da Daskalakis, Deckelbaum e Tzamos (2017), è che il meccanismo ottimale multi-dimensionale è in generale *non caratterizzabile in forma chiusa*: la sua struttura dipende in modo intricato dalla geometria della distribuzione dei tipi e non può essere espressa attraverso formule analitiche generalizzabili.

Il risultato di Crémer-McLean e la sua fragilità

Nonostante l'assenza di caratterizzazioni generali, il risultato di Crémer e McLean (1988) mostra qualcosa di sorprendente quando i tipi sono *correlati*: con tipi correlati in modo non banale, esistono meccanismi bayesianamente ottimali che estraggono l'intero surplus degli agenti in equilibrio. La correlazione tra i tipi consente al progettista di inferire il tipo privato di ogni agente dalle dichiarazioni degli altri, eliminando la "rendita per ignoranza" che caratterizza il caso con tipi indipendenti.

Il meccanismo di Crémer-McLean funziona attraverso scommesse: il progettista offre a ogni agente una scommessa sulla distribuzione dei tipi degli altri che è accettabile per ogni tipo dell'agente ma che, in aggregato,

consente al progettista di estrarre tutto il surplus. La scommessa è costruita sfruttando la correlazione tra i tipi.

La fragilità del risultato è però devastante per le applicazioni: il meccanismo funziona soltanto per la distribuzione esatta dei tipi specificata nel modello. Qualsiasi imprecisione nella stima di questa distribuzione può produrre meccanismi che non soddisfano la BIC o la IIR. Questa fragilità è particolarmente problematica per la governance dell'AI, dove l'incertezza sulla distribuzione dei tipi degli attori è profonda e strutturale — esattamente la motivazione per il meccanismo design robusto della sezione successiva.

II.8 — Meccanismo design robusto

Il problema del prior come common knowledge

Il meccanismo design classico assume che il progettista conosca la distribuzione dei tipi $p(\theta)$. Nella versione bayesiana, è addirittura richiesto che questa distribuzione sia common knowledge tra i giocatori. Nella pratica della governance dell'AI, questa conoscenza è raramente disponibile: il regolatore non conosce la distribuzione delle valorizzazioni private degli sviluppatori, la distribuzione dei loro profili di rischio, o la struttura di correlazione tra i tipi.

Robert Wilson (1987) articolò questa critica come obiezione di principio alle analisi standard — la *Wilson doctrine*: i meccanismi dovrebbero essere progettati per funzionare robustamente senza richiedere troppa common knowledge. L'obiezione non è di natura tecnica: riflette una preoccupazione fondamentale sulla fragilità dei meccanismi bayesiani

ottimali quando vengono messi alla prova da attori che non condividono il prior del progettista.

Equivalenza tra implementazione ex-post e DSIC

Il meccanismo design robusto, sviluppato sistematicamente da Bergemann e Morris (2005), formalizza l'intuizione di Wilson. La nozione centrale è l'*implementazione ex-post*: un meccanismo implementa f in modo ex-post se, per ogni profilo di tipi θ e qualsiasi sistema di credenze degli agenti compatibile con l'informazione comune, la dichiarazione veritiera del proprio tipo è un equilibrio bayesiano che produce l'esito $f(\theta)$.

Teorema (Bergemann e Morris, 2005): L'implementazione ex-post è equivalente all'implementazione in strategie dominanti: un meccanismo implementa f in modo ex-post per ogni possibile tipo space compatibile con le informazioni degli agenti se e solo se lo implementa in strategie dominanti.

Questo risultato rafforza enormemente la motivazione per la classe dei meccanismi DSIC: sono esattamente quelli che implementano f robustamente per qualsiasi distribuzione dei tipi e qualsiasi sistema di credenze degli agenti. Il prezzo della robustezza è l'efficienza — i meccanismi DSIC come il VCG sono in genere meno efficienti dei meccanismi bayesiani ottimali calibrati su una distribuzione specifica — ma per contesti come la governance dell'AI, dove l'incertezza sulla distribuzione dei tipi è strutturale, il guadagno in robustezza può più che compensare la perdita in efficienza.

La decisione tra meccanismi DSIC e meccanismi bayesiani ottimali è dunque, nella sua essenza, una decisione sul trade-off tra efficienza e

robustezza — e questa decisione dipende dalla fiducia che il progettista ripone nelle proprie stime del prior. Nella governance di tecnologie nuove e in rapida evoluzione, dove le distribuzioni dei tipi cambiano con le capacità tecnologiche, la robustezza ha un valore che spesso supera l'efficienza.

II.9 — Meccanismo design dinamico: impegno e time-inconsistency

Il principio di rivelazione dinamico

I meccanismi discussi finora sono prevalentemente statici. La realtà della governance è dinamica: le valorizzazioni cambiano nel tempo, nuova informazione si rivela nel corso dell'interazione, le capacità tecnologiche evolvono, e le strategie degli agenti si adattano in risposta alle politiche del regolatore.

Il *principio di rivelazione dinamico*, dovuto a Myerson (1986) e Hammond (1979), generalizza il principio statico: ogni meccanismo dinamico che implementa una funzione di allocazione dinamica in equilibrio bayesiano può essere sostituito da un meccanismo diretto in cui i tipi sono rivelati veritieramente in ogni periodo, condizionatamente alla storia passata. La struttura ricorsiva della rivelazione dinamica parallela quella dell'equazione di Bellman:

$$V_t(\theta_t, h_t) = \max_{\sigma_t} \{ u_t(\sigma_t, \theta_t) + \delta \cdot E_{\theta_{t+1}} \{ V_{t+1}(\theta_{t+1}, h_{t+1}) \} \}$$

dove h_t è la storia delle azioni e delle dichiarazioni fino al periodo t . Il meccanismo ottimale deve essere piano-compatibile in ogni punto della storia — le promesse fatte dal progettista devono essere mantenute per essere credibili.

Il ratchet effect e il problema del commitment

Il problema cruciale del meccanismo design dinamico è quello del *commitment*: il meccanismo produce gli incentivi desiderati soltanto se il progettista è credibilmente in grado di rispettare le regole del meccanismo anche quando, ex-post, avrebbe incentivo a deviarvi.

Il *ratchet effect*, analizzato da Freixas, Guesnerie e Tirole (1985) nel contesto della regolamentazione di un monopolio privato, illustra il problema con chiarezza. Il regolatore propone un contratto nel primo periodo, osserva il costo realizzato dell'impresa, e propone un nuovo contratto nel secondo periodo. Se l'impresa sa che un buon rendimento nel primo periodo porterà a condizioni più stringenti nel secondo — perché il regolatore usa l'informazione rivelata per estrarre più surplus — ha incentivo a limitare artificialmente la propria performance nel primo periodo, anche se questo è costoso per lei. Il meccanismo ottimale richiede che il regolatore si impegni a *non* usare le informazioni rivelate per modificare le condizioni future.

Formalmente, il ratchet effect emerge perché il progettista ha un incentivo temporale a sfruttare le informazioni rivelate. Sia θ_1 il tipo rivelato nel primo periodo. La scelta ottimale del progettista nel secondo periodo è:

$$\max_{x_2} u_P(x_2, \theta_1)$$

che dipende da θ_1 . L'agente, anticipando questa dipendenza, sceglie nel primo periodo non la strategia che rivela θ_1 , ma quella che rivela il θ_1 che massimizza le proprie utilità in entrambi i periodi. Il risultato è distorsione nell'equilibrio di entrambi i periodi.

Isomorfismo con la time-inconsistency macroeconomica

Il problema del commitment nel meccanismo design dinamico è strutturalmente isomorfo alla *time-inconsistency* della politica macroeconomica, identificata da Kydland e Prescott (1977). La credibilità del commitment a non inflazionare in futuro è necessaria per ancorare le aspettative inflazionistiche oggi, ma ex-post il banchiere centrale potrebbe avere incentivo a inflazionare per ridurre la disoccupazione. La soluzione istituzionale — l'indipendenza della banca centrale, che rende vincolante il commitment del governatore tramite il mandato istituzionale — è il meccanismo di commitment che rende credibile la politica monetaria.

Un meccanismo analogo serve per la governance dell'AI: l'indipendenza istituzionale del regolatore rispetto ai cicli politici e alle pressioni dell'industria rende credibile il commitment a non rinegoziare le regole in risposta a pressioni degli interessi industriali. Senza questo impegno di credibilità, gli sviluppatori razionali anticipano la rinegoziazione e rivelano strategicamente informazioni che portano a regole meno stringenti — esattamente il ratchet effect nel contesto della governance. Il problema non è la malafede del regolatore: è la struttura degli incentivi che la sua posizione crea, indipendentemente dalle intenzioni.

II.10 — La teoria del matching: stabilità senza prezzi

Il problema del matching

Una branca del meccanismo design di grande rilevanza applicativa è la teoria del matching, che studia come allocare agenti tra loro in modo stabile in *assenza di trasferimenti monetari*. Il problema differisce fondamentalmente dalle aste perché non prevede prezzi: la "valuta" dell'allocazione è la preferenza ordinale degli agenti. Questa differenza è decisiva in contesti dove i prezzi sono socialmente inaccettabili o istituzionalmente vietati — mercati dei reni, assegnazione dei medici negli ospedali, ammissione alle università.

Un *matching* $\mu : S \rightarrow C$ tra studenti e scuole è *stabile* se non esiste una coppia (studente s , scuola c) che si preferiscono reciprocamente rispetto alle proprie assegnazioni correnti nel senso che s preferisce c a $\mu(s)$ e c preferisce s a qualche studente che le è stato assegnato (o ha posti liberi). La stabilità cattura l'assenza di blocchi: nessuna coppia ha incentivo a formare un accordo bilaterale che aggirate il meccanismo. Questa nozione di stabilità è la controparte, nel matching senza prezzi, dell'equilibrio di Nash nel matching con prezzi.

L'algoritmo di Gale-Shapley

L'*algoritmo di Gale-Shapley* (o algoritmo dell'offerta differita, 1962) produce sempre un matching stabile in tempo finito. Nella versione studenti-propositori:

In ogni round, ogni studente non ancora assegnato fa un'offerta alla scuola più preferita tra quelle che non lo hanno ancora rifiutato. Ogni

scuola "trattiene" i migliori candidati fino alla propria capacità tra quelli che le hanno fatto offerta in questo round o che erano in lista d'attesa, e rifiuta tutti gli altri. Si itera fino a che tutti gli studenti sono assegnati o hanno esaurito la lista.

L'algoritmo termina in al più $|S| \times |C|$ passi — dove $|S|$ è il numero di studenti e $|C|$ il numero di scuole. Il matching prodotto soddisfa il teorema di ottimalità: è il matching stabile *Pareto-ottimale per gli studenti* — non esiste altro matching stabile che tutti gli studenti preferiscano. Al contempo, è il matching stabile *peggiore per le scuole*.

Il teorema di Roth (1982) mostra che il meccanismo studenti-propositori è *strategy-proof per gli studenti*: nessuno studente può ottenere un'assegnazione migliore falsificando la propria lista di preferenze. Questo parziale risultato di strategy-proofness — valido per il lato dei propositori ma non per il lato ricevente — è un risultato tipico dei meccanismi di matching: la manipolabilità da uno dei lati è intrinseca alla struttura del problema.

Un risultato importante di Roth e Sotomayor (1990) mostra che in ogni gioco di matching bipartito, l'insieme dei matching stabili forma un *reticolo* rispetto all'ordinamento dato dalle preferenze degli studenti: esiste un matching stabile ottimale per gli studenti (quello prodotto dall'algoritmo di Gale-Shapley con studenti propositori) e uno ottimale per le scuole (quello prodotto dall'algoritmo con scuole propositori), e tutti gli altri matching stabili sono "tra" questi due estremi. Questa struttura di reticolo è una proprietà geometrica profonda che non ha analogo nei matching con prezzi.

Applicazioni alla governance dell'AI

Per la governance dell'AI, la teoria del matching è rilevante in contesti dove i "prezzi" della certificazione non sono la valuta rilevante per l'allocazione ottimale.

Il primo contesto è il *matching tra valutatori e sistemi*: come assegnare team di valutazione indipendente ai sistemi di AI in modo da massimizzare la corrispondenza tra competenze e requisiti tecnici, evitando conflitti di interesse. I team di valutatori hanno preferenze sull'assegnazione (vogliono lavorare su sistemi che corrispondono alle proprie competenze), i sistemi hanno requisiti (necessitano di valutatori con specifiche competenze tecniche), e la stabilità del matching garantisce che nessuna coppia valutatore-sistema preferisca reciprocamente un'assegnazione diversa rispetto a quella prodotta dal meccanismo.

Il secondo contesto è il *matching tra usi approvati e sistemi*: quali sistemi di AI possono essere impiegati in quali contesti (sanità, finanza, sicurezza pubblica), rispettando vincoli di sicurezza e massimizzando il benessere degli utenti. La struttura coercitiva dei requisiti di sicurezza introduce vincoli aggiuntivi che un matching standard non cattura, ma le tecniche del matching con vincoli — sviluppate da Kamada e Kojima (2015) nel contesto del matching degli studenti con quote per regioni geografiche — possono essere estese a coprire questi casi.

II.11 — Verso la mappa delle tensioni strutturali

Le analisi di questa Parte hanno tracciato il perimetro del possibile nella progettazione delle istituzioni, rivelando un catalogo di tensioni che non

sono difetti contingenti dei meccanismi analizzati ma caratteristiche strutturali dello spazio dei meccanismi possibili.

Il meccanismo VCG è il vertice dell'architettura — l'unico meccanismo efficiente in strategie dominanti — ma non è budget-bilanciato. Il meccanismo ottimale di Myerson massimizza il ricavo ma sacrifica l'efficienza allocativa. Il meccanismo di Gale-Shapley produce stabilità senza prezzi ma è manipolabile dal lato non-propositori. L'implementazione robusta richiede l'implementazione in strategie dominanti, che a sua volta richiede il VCG, che non è budget-bilanciato. Il meccanismo di Crémer-McLean estrae l'intero surplus quando i tipi sono correlati, ma è fragile a qualsiasi deviazione dal prior corretto.

Il teorema di Myerson-Satterthwaite formalizza il più profondo di questi trade-off: efficienza, incentivi alla rivelazione, partecipazione volontaria e auto-finanziamento non sono simultaneamente raggiungibili in nessun meccanismo di contrattazione bilaterale con tipi privati sovrapposti. Il teorema di Green-Laffont mostra che la classe VCG è l'unica che combina efficienza e strategie dominanti. Il teorema di Bergemann-Morris mostra che la robustezza richiede strategie dominanti.

Queste impossibilità non sono la fine del discorso sulla progettazione istituzionale: sono il suo inizio. Conoscere con precisione cosa non si può fare è la condizione per scegliere consapevolmente cosa fare — quale compromesso privilegiare, quale proprietà sacrificare, quale dimensione del problema merita la priorità in ciascun contesto specifico. Il progettista di istituzioni che non conosce questa mappa è condannato a scoprire i limiti del possibile nel modo peggiore: costruendo meccanismi che sembrano funzionare ma crollano quando vengono messi alla prova dalla razionalità strategica degli agenti reali.

TEORIA DEI GIOCHI EVOLUTIVI

III.1 — Il passaggio dalla razionalità all'evoluzione

La critica alle assunzioni cognitive

La teoria dei giochi classica è una teoria della razionalità individuale: agenti che massimizzano l'utilità attesa, formano credenze bayesianamente coerenti, calcolano equilibri attraverso ragionamento introspettivo deliberato. Questa architettura ha una potenza analitica straordinaria — ma le sue assunzioni cognitive sono raramente soddisfatte pienamente da agenti reali.

Gli individui commettono violazioni sistematiche delle assunzioni di razionalità, ampiamente documentate dalla psicologia economica sperimentale inaugurata da Kahneman e Tversky (1979) con la loro teoria del prospetto: sovrastimano le piccole probabilità, sottostimano le grandi, invertono le preferenze in funzione del framing, sono affetti da bias di status quo, di ancoraggio, di eccesso di confidenza, di aversione alle perdite. Le organizzazioni non massimizzano funzioni di utilità condivise ma sono arene di conflitti interni, di routine cristallizzate, di procedure decisionali che riflettono la storia più che la razionalità. Gli algoritmi di apprendimento automatico — inclusi i sistemi AI la cui governance è l'oggetto di questo saggio — non calcolano equilibri in forma chiusa: ottimizzano funzioni obiettivo attraverso gradient descent su dati empirici, e i loro "comportamenti strategici" emergono da questo processo in modi che spesso sorprendono i loro stessi progettisti.

La teoria dei giochi evolutivi, inaugurata da Maynard Smith e Price (1973) con il concetto di strategia evolutivamente stabile, e sistematizzata da Weibull (1995) e da Hofbauer e Sigmund (1998), compie un passaggio concettuale fondamentale. La domanda non è più "quale strategia un singolo agente razionale *sceglierebbe?*" ma "quali strategie si *diffonderanno* in una popolazione di agenti che si adattano attraverso selezione, imitazione o apprendimento?" L'analogia con la biologia darwiniana è strutturalmente precisa: in biologia, la frequenza di un genotipo cresce se gli individui che lo portano hanno fitness media superiore alla media della popolazione; nella teoria dei giochi evolutivi, la frequenza di una strategia cresce se i suoi portatori ottengono payoff medi superiori alla media. La matematica è identica; l'interpretazione è diversa.

Il duplice contributo della prospettiva evolutiva

Il contributo specifico della teoria evolutiva che la distingue dall'approccio classico è duplice, e vale la pena articolarlo con precisione prima di procedere alle derivazioni formali.

Il primo contributo è l'eliminazione del requisito di razionalità consapevole. Le dinamiche evolutive funzionano per *qualsiasi* meccanismo di adattamento — selezione naturale, imitazione dei vicini più performanti, apprendimento per rinforzo con aggiornamenti locali, algoritmi genetici — purché il meccanismo soddisfi la condizione che le strategie più performanti si diffondano a spese delle meno performanti. L'assenza di calcolo deliberato è un pregio, non una limitazione: rende la teoria applicabile a una gamma enormemente più ampia di sistemi adattativi, inclusi sistemi AI che apprendono, organizzazioni che adottano pratiche di settore, stati che imitano le politiche di altri stati con risultati migliori.

Il secondo contributo è lo studio della *stabilità dinamica* degli equilibri, in contrasto con la loro mera esistenza statica. Un equilibrio di Nash può essere staticamente stabile — nessun giocatore ha incentivo a deviare unilateralmente — ma dinamicamente instabile: una piccola perturbazione nella distribuzione delle strategie della popolazione porta il sistema ad allontanarsi dall'equilibrio invece di tornarvi. La teoria evolutiva classifica gli equilibri in attrattori stabili, punti sella instabili e cicli limite, rivelando quale parte dello spazio degli equilibri di Nash è effettivamente raggiungibile attraverso processi di adattamento naturale — e quale è un artefatto del formalismo statico.

III.2 — Le dinamiche di replica: assiomatizzazione e proprietà

Derivazione assiomatica

Consideriamo una popolazione con vettore di frequenze $p = (p_1, \dots, p_n) \in \Delta^n$, dove p_i è la frazione di agenti che adotta la strategia s_i , e Δ^n è il simpleso standard. Con matrice di payoff A , il payoff medio della strategia s_i è:

$$f_i(p) = (Ap)_i = \sum_j A_{ij} p_j$$

e il payoff medio della popolazione è:

$$f(p) = p^T Ap = \sum_i p_i f_i(p)$$

Nachbar (1990) dimostrò che tre condizioni assiomatiche naturali determinano univocamente la forma funzionale delle dinamiche di adattamento:

Invarianza del semplice: $\sum_i \dot{p}_i = 0$, le frequenze rimangono una distribuzione di probabilità per ogni $t \geq 0$.

Positività della selezione: $f_i > f_j$ implica $\dot{p}_i/p_i > \dot{p}_j/p_j$, le strategie più performanti crescono più rapidamente in termini relativi.

Rispetto dei bordi: $p_i(0) = 0$ implica $p_i(t) = 0$ per ogni $t \geq 0$, le strategie assenti restano assenti — non emerge spontaneamente variazione genetica nuova.

L'unica forma funzionale continua compatibile con queste tre condizioni è le *dinamiche di replica*:

$$\dot{p}_i = p_i \cdot [f_i(p) - f(p)]$$

La strategia i cresce se e solo se il suo payoff supera la media della popolazione, e decresce altrimenti. La velocità di crescita relativa $\dot{p}_i / p_i = f_i - f$ è proporzionale alla deviazione del payoff dalla media.

Queste dinamiche ammettono un'interpretazione diretta come processo di imitazione: in ogni istante, ogni agente osserva un altro agente scelto a caso, e con probabilità proporzionale alla differenza positiva di payoff imita la sua strategia. Il risultato aggregato di questo processo di imitazione locale è precisamente le dinamiche di replica — un risultato di Weibull (1995) che connette la microfondazione comportamentale alla dinamica aggregata.

La legge di Fisher

La proprietà fondamentale delle dinamiche di replica è la *crescita monotona del payoff medio* — la legge di Fisher, che nella biologia evolutiva è conosciuta come il Teorema Fondamentale della Selezione Naturale (Fisher, 1930):

$$\begin{aligned} df/dt &= d(p^T A p)/dt = \sum_i (\partial f / \partial p_i) \cdot \dot{p}_i \\ &= \sum_i 2(Ap)_i \cdot p_i(f_i - \bar{f}) \\ &= 2 \sum_i p_i f_i (f_i - \bar{f}) \\ &= 2 [\sum_i p_i f_i^2 - \bar{f}^2] \\ &= 2 \text{Var}(f) \geq 0 \end{aligned}$$

dove $\text{Var}(f) = \sum_i p_i (f_i - \bar{f})^2$ è la varianza del payoff nella popolazione. Il payoff medio è monotonicamente non-decrescente lungo ogni traiettoria, con velocità di crescita proporzionale alla varianza del payoff. I punti fissi sono gli stati in cui tutte le strategie presenti hanno lo stesso payoff medio — condizione che coincide con la definizione di equilibrio di Nash per il gioco simmetrico.

Questa legge ha implicazioni immediate: la varianza del payoff è il "carburante" dell'evoluzione. Popolazioni con alta varianza evolvono rapidamente verso payoff più alti; popolazioni con bassa varianza evolvono lentamente. Un ecosistema di sviluppatori AI in cui le pratiche di sicurezza variano molto tra organizzazioni — alta varianza nelle pratiche — evolverà più rapidamente di uno in cui tutti adottano le stesse pratiche, nel bene e nel male.

Il teorema di classificazione

Il teorema di classificazione fondamentale (Zeeman 1980; Hofbauer e Sigmund 1998) stabilisce le relazioni tra le dinamiche di replica e i concetti di equilibrio statici:

Prima parte: ogni equilibrio di Nash *asintoticamente stabile* delle dinamiche di replica è un equilibrio di Nash del gioco.

Seconda parte: ogni equilibrio di Nash *strettamente dominante* — dove una strategia domina strettamente tutte le altre per ogni giocatore — è asintoticamente stabile.

Terza parte: le strategie *strettamente dominate* sono eliminate nel lungo periodo dalla selezione — replicando in forma dinamica il risultato statico dell'eliminazione iterata delle strategie dominate.

La prima parte dice che la selezione non può produrre stati che non siano equilibri di Nash — la stabilità evolutiva implica la stabilità di Nash. La seconda parte dice che gli equilibri con dominio stretto sono evolutivamente inevitabili — se un agente ha una strategia che fa meglio in ogni circostanza, la selezione la porterà alla dominanza. La terza parte è il risultato dinamico più utile: elimina sistematicamente le strategie che non sono mai le migliori risposta, purificando il simplesso delle strategie "irrazionali".

Importante: non ogni equilibrio di Nash è un punto fisso stabile. Gli equilibri che dipendono da strategie strettamente dominate — inclusi molti degli equilibri con minacce non credibili discussi nella Parte Prima — sono instabili sotto le dinamiche di replica: piccole perturbazioni portano il sistema ad allontanarsi da essi. L'instabilità evolutiva è il meccanismo che depura lo spazio degli equilibri dalla proliferazione di soluzioni non robuste.

III.3 — La strategia evolutivamente stabile

Definizione e condizioni

Una strategia σ^* è *evolutivamente stabile* (ESS, Evolutionarily Stable Strategy) se resiste all'invasione di qualsiasi mutante: per ogni strategia alternativa $\sigma \neq \sigma^*$, esiste $\bar{\epsilon} > 0$ tale che per ogni $\epsilon \in (0, \bar{\epsilon})$, la strategia σ^* ottiene payoff maggiore di σ in una popolazione $(1-\epsilon)\sigma^* + \epsilon\sigma$:

$$U(\sigma^*, (1-\epsilon)\sigma^* + \epsilon\sigma) > U(\sigma, (1-\epsilon)\sigma^* + \epsilon\sigma)$$

La condizione si decompone in due casi distinti che rivelano due meccanismi di stabilità diversi, e vale la pena esplicitarli separatamente perché hanno interpretazioni qualitativamente diverse.

Caso I — stabilità di primo ordine: $U(\sigma^*, \sigma^*) > U(\sigma, \sigma^*)$ per ogni $\sigma \neq \sigma^*$. La strategia residente domina i mutanti nell'interazione con un residente — la selezione di primo ordine è sufficiente. Basta che i portatori della strategia residente ottengano payoff superiori ai mutanti nell'interazione con la maggioranza della popolazione (che è residente) per garantire la stabilità.

Caso II — stabilità di secondo ordine: $U(\sigma^*, \sigma^*) = U(\sigma, \sigma^*)$ ma $U(\sigma^*, \sigma) > U(\sigma, \sigma)$ per ogni $\sigma \neq \sigma^*$. La parità nel confronto principale è superata da un vantaggio nel confronto incrociato — la strategia residente fa meglio del mutante nell'interazione con il mutante stesso. La selezione di secondo ordine è necessaria quando la selezione di primo ordine è neutrale.

Il Caso II è particolarmente rilevante per la governance: illustra come strategie che sembrano equivalenti nel confronto principale possano differire nella loro robustezza all'invasione di mutanti. In un ecosistema di sviluppatori dove le pratiche di sicurezza hanno il medesimo payoff atteso medio, la strategia più robusta evolutivamente è quella che fa meglio di qualsiasi alternativa nell'interazione con quell'alternativa stessa.

Relazioni con l'equilibrio di Nash

Teorema (Maynard Smith, 1974): Ogni ESS è un equilibrio di Nash simmetrico del gioco. Non ogni equilibrio di Nash simmetrico è ESS.

Dimostrazione della prima parte. Supponiamo per assurdo che σ^* sia ESS ma non equilibrio di Nash. Allora esiste σ tale che $U(\sigma, \sigma^*) > U(\sigma^*, \sigma^*)$. Per continuità di U , per ϵ sufficientemente piccolo, $U(\sigma, (1-\epsilon)\sigma^* + \epsilon\sigma) > U(\sigma^*, (1-\epsilon)\sigma^* + \epsilon\sigma)$ — contraddicendo la definizione di ESS. $\square$

Ogni ESS è anche un attrattore asintoticamente stabile delle dinamiche di replica. La relazione inversa non vale: esistono attrattori ciclici — cicli limite — nei giochi con tre o più strategie che non corrispondono a nessun ESS. Il caso prototipico è il gioco roccia-carta-forbici: l'unico equilibrio simmetrico è il punto fisso interno $(1/3, 1/3, 1/3)$, ma questo equilibrio è circondato da cicli neutrali — le traiettorie delle dinamiche di replica girano attorno ad esso senza convergervi né divergere. Questa dinamica oscillatoria ha analoghi diretti nei sistemi AI addestrati in ambienti multi-agente, come si vedrà nella sezione III.8.

Implicazioni per la governance

La distinzione tra ESS e equilibrio di Nash ha implicazioni immediate per la governance dell'AI. Un regime di governance che è un equilibrio di Nash — nessun singolo attore ha incentivo a deviare unilateralmente —

ma non è evolutivamente stabile può essere eroso nel tempo da pratiche "mutanti": organizzazioni che adottano comportamenti devianti e ottengono vantaggi competitivi, spingendo altri ad imitarli attraverso il meccanismo dell'imitazione dei vincitori. Solo i regimi di governance che sono evolutivamente stabili — nel senso che le pratiche di conformità ottengono payoff superiori alle pratiche devianti anche quando queste ultime sono rare — sono robustamente sostenibili nel lungo periodo senza enforcement continuo e costoso.

III.4 — Dinamiche stocastiche: il processo di Moran e la probabilità di fissazione

Le dinamiche in popolazioni finite

Le dinamiche di replica sono dinamiche deterministiche appropriate per popolazioni infinite o molto numerose. In popolazioni di dimensione finita N — come l'ecosistema effettivo degli sviluppatori AI, che conta alcune decine di organizzazioni di rilievo — le dinamiche evolutive sono *stocastiche*: le fluttuazioni demografiche casuali (la deriva genetica della biologia evolutiva) possono portare all'estinzione di strategie evolutivamente favorite o alla fissazione di strategie sfavorevoli.

Il *processo di Moran* modella questa stocasticità in modo minimale. In ciascun passo discreto: un agente è selezionato per la "riproduzione" con probabilità proporzionale alla propria fitness relativa; sostituisce un agente scelto uniformemente a caso. Questo processo è una catena di Markov con stati $\{0, 1, \dots, N\}$ (numero di cooperatori) e due stati assorbenti: 0 (tutti defezionanti) e N (tutti cooperatori).

Con k cooperatori su N agenti totali, le probabilità di transizione sono:

$$T^+(k) = [k \cdot f_C(k) / (k \cdot f_C(k) + (N-k) \cdot f_D(k))] \cdot (N-k)/N$$

$$T^-(k) = [(N-k) \cdot f_D(k) / (k \cdot f_C(k) + (N-k) \cdot f_D(k))] \cdot k/N$$

dove $f_C(k)$ e $f_D(k)$ sono le fitness di cooperatori e defezionanti rispettivamente. La probabilità di transizione T^+ è il prodotto della probabilità di selezionare un cooperatore per la riproduzione (primo fattore, proporzionale alla fitness relativa) e della probabilità che l'agente sostituito sia un defezionante (secondo fattore, uniforme).

La probabilità di fissazione: derivazione

La *probabilità di fissazione* q_k — la probabilità di raggiungere lo stato $k=N$ partendo dallo stato k — soddisfa il sistema di equazioni alle differenze che esprime la media ponderata delle probabilità di fissazione degli stati adiacenti:

$$q_k = T^+(k) q_{k+1} + T^0(k) q_k + T^-(k) q_{k-1}$$

con condizioni al contorno $q_0 = 0$ e $q_N = 1$. Riscriviamo come:

$$T^+(k)(q_{k+1} - q_k) = T^-(k)(q_k - q_{k-1})$$

Definendo $\Delta_k = q_k - q_{k-1}$ e $\gamma_k = T^-(k)/T^+(k)$, si ha:

$$\Delta_{k+1} = \gamma_k \Delta_k = \Delta_1 \cdot \prod_{j=1}^k \gamma_j$$

La condizione $q_N = 1$ dà $\sum_{k=1}^N \Delta_k = 1$, da cui:

$$q_1 = 1 / (1 + \sum_{k=1}^{N-1} \prod_{j=1}^k \gamma_j)$$

e più in generale:

$$q_m = \frac{\sum_{k=0}^{m-1} \prod_{j=1}^k \gamma_j}{\sum_{k=0}^{N-1} \prod_{j=1}^k \gamma_j}$$

La quantità $\gamma_k = T^-/T^+ = f_D(k)/f_C(k)$ è il rapporto di fitness tra defezionanti e cooperatori allo stato k . Se $\gamma_k < 1$ per ogni k , i cooperatori hanno fitness superiore in ogni stato e la probabilità di fissazione cresce rapidamente con N . Se $\gamma_k > 1$, i defezionanti hanno il vantaggio e la cooperazione è evolutivamente svantaggiata.

Condizione di neutralità: In assenza di selezione (tutte le fitness uguali a 1), $\gamma_k = 1$ per ogni k , e:

$$q_1 = 1 / (1 + (N-1)) = 1/N$$

La selezione favorisce il cooperatore rispetto alla deriva neutra se $q_1 > 1/N$.

La condizione $\sigma \geq N$ e l'intuizione del weak selection

Per *selezione debole* — il caso in cui le differenze di fitness sono piccole rispetto alla variazione demografica, formalmente $w \rightarrow 0$ con $f_i(k) = 1 - w + w \cdot \pi_i(k)$ — si può espandere γ_k al primo ordine in w e ottenere condizioni analitiche esplicite. Per il dilemma del prigioniero con parametri b (beneficio per il partner) e c (costo per il cooperatore), su grafo completo:

La condizione $q_1 > 1/N$ nella selezione debole diventa:

$$f_C(k) > f_D(k) \text{ in media su tutti gli stati } k \in \{1, \dots, N-1\}$$

che si traduce, per il dilemma del prigioniero su grafo completo, nella condizione:

$b/c > N/2 - 1 \approx N/2$ per N grande

Per popolazioni grandi, la cooperazione richiede che il beneficio superi il costo di un fattore proporzionale alla dimensione della popolazione. Questo spiega l'ubiqua difficoltà della cooperazione in contesti anonmi: con $N = 100$ sviluppatori, il beneficio della sicurezza dell'ecosistema dovrebbe essere 50 volte il costo della salvaguardia per un singolo sviluppatore, affinché la cooperazione sia evolutivamente favorita senza struttura di rete.

III.5 — Struttura della rete e la condizione $b/c > k$

Il ruolo della topologia

Le dinamiche di replica standard e il processo di Moran assumono popolazioni *well-mixed*: ogni agente interagisce con ogni altro con uguale probabilità. Questa assunzione è raramente soddisfatta nelle interazioni sociali ed economiche reali, dove la struttura delle relazioni — chi interagisce con chi, con quale intensità — modella profondamente le dinamiche evolutive.

Nowak e May (1992) dimostrarono, attraverso simulazioni su griglia che aprirono un filone di ricerca straordinariamente fecondo, che la struttura spaziale può sostenere la cooperazione in condizioni che porterebbero alla defezione in una popolazione *well-mixed*. Il meccanismo è la *segregazione spaziale dei cooperatori*: su una griglia, i cooperatori tendono a formare cluster che si proteggono reciprocamente dai defezionanti ai

margini. Il cluster cooperativo ottiene un payoff interno elevato (cooperatori che interagiscono con cooperatori), anche se paga un costo ai margini (cooperatori che interagiscono con defezionanti).

La condizione $b/c > k$: derivazione per pair approximation

Su reti regolari con k vicini per nodo, Ohtsuki, Hauert, Lieberman e Nowak (2006) derivarono, attraverso la *pair approximation*, la condizione necessaria e sufficiente per la cooperazione:

$$b/c > k$$

La pair approximation è un metodo di chiusura di momento: invece di tracciare le frequenze dei singoli agenti (approssimazione di campo medio), considera le frequenze delle *coppie* di agenti adiacenti $p_{\{CC\}}$, $p_{\{CD\}}$, $p_{\{DD\}}$, e approssima le frequenze delle triplette come funzioni delle coppie:

$$p_{\{XYZ\}} \approx p_{\{XY\}} \cdot p_{\{YZ\}} / p_Y$$

Questa chiusura è esatta in alcune topologie e approssimata in altre; la sua precisione migliora per reti con cicli lunghi.

Il sistema di equazioni differenziali per le frequenze delle coppie — derivato dalle transizioni del processo di Moran su rete — è:

$$\dot{p}_{\{CC\}} = 2 p_{\{CC\}} (f_{\{C|CC\}} - \bar{f}) + \text{termini di diffusione}$$

dove $f_{\{C|CC\}}$ è il payoff atteso di un cooperatore dato che il suo vicino è un cooperatore. Risolvendo il sistema per la condizione di stabilità del punto fisso $p_C = 0$ (tutti defezionanti) si ottiene, nell'approssimazione di selezione debole:

la condizione critica $b/c = k$

Al di sotto della soglia ($b/c < k$), i defezionanti invadono i cooperatori e la cooperazione non è evolutivamente stabile. Al di sopra della soglia ($b/c > k$), i cooperatori resistono all'invasione dei defezionanti e la cooperazione è evolutivamente favorita.

Il caso limite $k \rightarrow \infty$ corrisponde alla popolazione well-mixed, dove la condizione diventa $b/c > \infty$ — la cooperazione è impossibile nel dilemma del prigioniero well-mixed, coerentemente con i risultati classici. Il caso $k = 2$ (catena lineare) richiede soltanto $b/c > 2$ — la struttura lineare della rete è molto favorevole alla cooperazione.

Strutture di rete e implicazioni per la governance

La condizione $b/c > k$ ha implicazioni concrete per la progettazione delle strutture di governance dell'AI. L'ecosistema degli sviluppatori AI non è una popolazione well-mixed: è una rete di organizzazioni con relazioni di partnership, competizione, accesso a investitori comuni, collaborazioni di ricerca. La topologia di questa rete è un parametro di governance — non un fatto contingente dell'organizzazione industriale.

Ecosistemi strutturati come reti sparse di cluster interconnessi — consorzi industriali su specifici domini applicativi, comunità di ricerca aperta con standard condivisi, accordi bilaterali tra laboratori per la condivisione di pratiche di sicurezza — possono sostenere la cooperazione nelle pratiche di sicurezza a rapporti b/c inferiori rispetto a ecosistemi con interazione diffusa e indifferenziata. Creare queste strutture di rete — attraverso incentivi, standard settoriali, infrastrutture di comunicazione e di condivisione dell'informazione — è un atto di governance con effetti evolutivi profondi e duraturi.

Esiste però un caveat importante: la topologia di rete che favorisce la cooperazione in un dilemma del prigioniero potrebbe non essere la stessa che favorisce la diffusione dell'innovazione o la resilienza a shock. Reti molto sparse con molti cluster separati sono ottime per la cooperazione ma lente nella diffusione di buone pratiche tra cluster. Reti dense con alta connettività sono ottime per la diffusione ma difficili per la cooperazione. Il design ottimale della struttura di rete per la governance dell'AI deve bilanciare questi obiettivi multipli — un problema di ottimizzazione multi-obiettivo per il quale non esiste soluzione universale.

III.6 — Selezione multi-livello e l'equazione di Price

L'equazione di Price

L'*equazione di Price* (Price, 1970) è il risultato più generale dell'intera teoria della selezione. Descrive il cambiamento nella frequenza media di qualsiasi carattere sotto qualsiasi forma di selezione a qualsiasi livello gerarchico, con un'eleganza matematica che pochi risultati nella scienza possono eguagliare.

Sia la popolazione suddivisa in gruppi $k = 1, \dots, G$. Sia p_k la frequenza del carattere (ad esempio, la frequenza dei cooperatori) nel gruppo k , w_k la fitness del gruppo k , e $\bar{w} = \sum_k (n_k/N) w_k$ la fitness media pesata per la dimensione del gruppo. L'equazione di Price è:

$$\bar{w} \cdot \Delta \bar{p} = \text{Cov}(w_k, p_k) + E[w_k \cdot \Delta p_k]$$

dove la covarianza e il valore atteso sono calcolati sulla distribuzione delle dimensioni dei gruppi.

Il primo termine — la covarianza tra fitness del gruppo e frequenza del carattere — cattura la *selezione tra gruppi*: è positiva se i gruppi con più cooperatori (p_k alto) hanno fitness più alta (w_k alto). Il secondo termine cattura la *selezione dentro i gruppi*: è tipicamente negativa nel dilemma del prigioniero, perché all'interno di ogni gruppo i defezionanti ottengono payoff maggiori dei cooperatori.

La cooperazione evolve se il primo termine supera il secondo in valore assoluto — se la selezione tra gruppi supera la selezione dentro i gruppi. Questa è la condizione formale per la *selezione di gruppo*: la cooperazione può evolvere non perché i cooperatori abbiano vantaggi individuali, ma perché i *gruppi* di cooperatori si riproducono più rapidamente dei gruppi misti o di defezionanti puri.

La regola di Hamilton come caso speciale

La *regola di Hamilton* $rb > c$ emerge come caso speciale dell'equazione di Price. Riscrivendo la covarianza in termini del coefficiente di parentela r — la probabilità che, selezionando casualmente un individuo dello stesso gruppo, esso condivida il carattere del donatore — si ottiene:

$$\text{Cov}(w_k, p_k) \approx r \cdot b \cdot \text{Var}(p)$$

dove b è il beneficio della cooperazione per il partner e $\text{Var}(p)$ è la varianza inter-gruppo della frequenza del carattere. La condizione di evoluzione della cooperazione diventa:

$$r \cdot b \cdot \text{Var}(p) > c \cdot \text{Var}(p) \Leftrightarrow rb > c$$

Il coefficiente r non deve essere interpretato necessariamente come parentela genetica: è un coefficiente di correlazione che misura la probabilità di interagire con un individuo simile. Nella governance dell'AI, r è alto quando gli sviluppatori interagiscono prevalentemente con altri sviluppatori che condividono le proprie pratiche (alta assortitività della rete), e basso quando le interazioni sono casuali. La governance che crea strutture di interazione assortative — favorendo le interazioni tra organizzazioni con pratiche simili — aumenta r e quindi facilita l'evoluzione della cooperazione.

Implicazioni per la governance attraverso la struttura dei cluster

La governance dell'AI può sfruttare la logica della selezione multi-livello: *cluster di organizzazioni con elevati standard di sicurezza* — certificati, premiati nell'accesso agli appalti pubblici, strutturati come consorzi industriali con accesso preferenziale a risorse condivise — possono sopravvivere evolutivamente anche quando la selezione individuale penalizzi i costi delle salvaguardie, purché la selezione *tra cluster* sia sufficientemente forte.

La forza della selezione tra cluster dipende dalla struttura degli incentivi che il regolatore crea: differenziali di accesso al mercato, premi per la partecipazione a consorzi di sicurezza, penalità per incidenti che riducono la reputazione dell'intero cluster — tutti questi meccanismi aumentano la fitness dei cluster ad alto standard rispetto a quelli a basso standard, creando pressione selettiva che può sopravanzare la pressione competitiva individuale.

III.7 — Reciprocità indiretta e infrastrutture di reputazione

Il meccanismo della reputazione

La *reciprocità indiretta* (Nowak e Sigmund, 1998) opera attraverso un circuito reputazionale: agisco generosamente con te non perché ti conosca direttamente o perché mi aspetti una tua reciprocità futura, ma perché altri osservano la mia generosità e mi trattano bene in futuro sulla base di essa. Il sistema richiede due componenti: una *norma di valutazione* che assegni reputazioni agli individui in base al loro comportamento osservato, e una *norma di azione* che specifichi il comportamento condizionato alle reputazioni.

L'evoluzione della reciprocità indiretta richiede che l'informazione sulla reputazione sia disponibile e affidabile. In popolazioni piccole con interazioni ripetute frequenti, la reputazione può essere mantenuta attraverso osservazione diretta e gossip. In popolazioni grandi con interazioni rare tra coppie specifiche — come l'ecosistema globale degli sviluppatori AI — richiede *infrastrutture di reputazione* esplicite: database di incidenti, sistemi di certificazione pubblici, valutazioni di terze parti, obblighi di disclosure.

La condizione di evoluzione della cooperazione attraverso reciprocità indiretta è approssimativamente:

$$b \cdot q > c$$

dove q è la probabilità che il donante venga osservato e valutato correttamente durante l'interazione. Se q è alto — se le informazioni reputazionali sono accurate e ampiamente disponibili — la cooperazione

è evolutivamente favorita anche in popolazioni grandi e anonime. Se q è basso — se le informazioni sono scarse, inaccurate o accessibili solo a pochi — la reciprocità indiretta collassa e la cooperazione non può essere sostenuta.

Le Leading Eight

Ohtsuki e Iwasa (2004) hanno dimostrato che tra tutte le possibili norme di valutazione di secondo ordine — che considerano sia l'azione del donante sia la reputazione del ricevente — soltanto otto sono *evolutivamente stabili* in un contesto di population dynamics: le *Leading Eight*. Queste norme condividono la proprietà che assegnano buona reputazione alla cooperazione con i "buoni" e al rifiuto dei "cattivi", e cattiva reputazione alla cooperazione con i "cattivi" o al rifiuto dei "buoni".

La norma *Stern Judging*, una delle Leading Eight, funziona come segue: un donante riceve buona reputazione se coopera con un ricevente di buona reputazione o rifiuta di cooperare con uno di cattiva reputazione; riceve cattiva reputazione se coopera con un ricevente di cattiva reputazione o rifiuta di cooperare con uno di buona reputazione. La condizione di evoluzione della cooperazione sotto Stern Judging è $b \cdot q_G > c$, dove q_G è la probabilità che un cooperatore incontri un ricevente con buona reputazione.

L'interesse delle Leading Eight per la governance non è puramente teorico. Esse suggeriscono che le infrastrutture di reputazione efficaci devono incorporare norme di *giudizio contestuale* — non basta registrare se un'organizzazione ha cooperato o meno, ma occorre valutare con chi cooperava. Un'organizzazione che coopera con un'altra organizzazione nota per pratiche rischiose non merita buona reputazione per questa

cooperazione: secondo le Leading Eight, questa è anzi una ragione per assegnarle cattiva reputazione. Le infrastrutture di reputazione per l'AI che distinguono la qualità del partner — non solo la frequenza della cooperazione — sono quindi evolutivamente superiori a quelle che si limitano a registrare la frequenza delle azioni "cooperative" senza contesto.

Progettare le infrastrutture di reputazione

Le infrastrutture di reputazione per l'ecosistema AI — valutazioni pubbliche dei sistemi, certificazioni di sicurezza visibili, database di incidenti accessibili ai ricercatori, obblighi di trasparenza sulle pratiche di red-teaming e sulle vulnerabilità scoperte — sono la traduzione istituzionale della reciprocità indiretta. Non sono soltanto strumenti informativi: sono meccanismi evolutivi che possono sostenere la cooperazione nelle pratiche di sicurezza anche in assenza di interazioni dirette ripetute tra ogni coppia di attori nell'ecosistema.

Tre proprietà delle infrastrutture di reputazione efficaci emergono dall'analisi delle Leading Eight.

Accuratezza: le valutazioni devono essere sufficientemente precise da permettere la discriminazione tra tipi. Un database di incidenti che non distingue tra violazioni gravi e minori, o che non tiene conto del contesto in cui si è verificato l'incidente, non fornisce il segnale reputazionale necessario per sostenere la cooperazione. L'accuratezza richiede protocolli di valutazione standardizzati, validati e verificabili da parti terze.

Accessibilità: le informazioni di reputazione devono essere disponibili agli attori che le usano per prendere decisioni — deployer che scelgono quali

sistemi usare, investitori che allocano capitale, ricercatori che valutano la qualità delle pratiche di sicurezza, altri sviluppatori che decidono con chi collaborare. L'accessibilità non è automatica: richiede infrastrutture tecniche e legali per la condivisione delle informazioni che superino le resistenze degli attori che preferirebbero mantenere opacità sulle proprie pratiche.

Immutabilità: le valutazioni passate devono essere resistenti alla revisione retrospettiva da parte degli attori che le ricevono. Un sistema in cui le organizzazioni possono fare pressione per rimuovere le valutazioni negative dalla loro storia reputazionale perde il suo valore evolutivo, perché elimina il segnale che distingue i tipi nel lungo periodo. L'immutabilità è una proprietà istituzionale che richiede indipendenza dell'ente gestore dell'infrastruttura di reputazione dagli attori che valuta.

III.8 — Apprendimento multi-agente e instabilità strutturale del training

Il problema del training in ambienti non stazionari

Il training di sistemi AI in ambienti multi-agente affronta un problema fondamentale che la teoria dei giochi evolutivi illumina in modo diretto. Consideriamo due agenti che si addestrano simultaneamente, ciascuno con parametri θ_1 e θ_2 che aggiornano attraverso gradient ascent sui propri obiettivi $u_1(\theta_1, \theta_2)$ e $u_2(\theta_1, \theta_2)$:

$$\theta_1 = \nabla \{ \theta_1 \} u_1(\theta_1, \theta_2) \quad \theta_2 = \nabla \{ \theta_2 \} u_2(\theta_1, \theta_2)$$

In un gioco a somma zero ($u_1 = -u_2 = u$), il sistema si riduce a:

$$\theta_1 = \nabla \{\theta_1\} u \quad \theta_2 = -\nabla \{\theta_2\} u$$

Il campo vettoriale di questo sistema si scompone in due componenti ortogonali. Definiamo $\xi = (\theta_1, \theta_2) \in \mathbb{R}^{d_1+d_2}$ e scriviamo il campo vettoriale come $v(\xi) = (\nabla \{\theta_1\} u, -\nabla \{\theta_2\} u)$. Seguendo Balduzzi e Tuyls (2018), si scompone:

$$v(\xi) = \nabla V(\xi) + R(\xi)$$

dove la componente *simmetrica* $\nabla V(\xi) = (\nabla \{\theta_1\} u + \nabla \{\theta_2\}^T J \cdot \nabla \{\theta_1\} u, \dots)$ porta le traiettorie verso i punti stazionari, e la componente *antisimmetrica* $R(\xi)$ — il campo rotazionale — causa orbite attorno ai punti di equilibrio invece della convergenza.

Più precisamente, per un campo vettoriale $v(\xi)$, la parte simmetrica della matrice Jacobiana $J^s = (J + J^T)/2$ determina la dinamica di convergenza, mentre la parte antisimmetrica $J^a = (J - J^T)/2$ determina la dinamica rotazionale. Un punto fisso è un attrattore se e solo se tutti gli autovalori di J^s sono negativi. Per i giochi a somma zero, la parte simmetrica di J è zero ai punti di equilibrio — rendendo l'analisi della stabilità impossibile tramite la sola linearizzazione.

Il Consensus Optimization

Balduzzi e Tuyls propongono il *Consensus Optimization* come correzione che riduce la componente rotazionale aggiungendo un termine di penalty che attrae le traiettorie verso regioni di basso gradiente:

$$\theta_1 = \nabla \{\theta_1\} u_1 + \lambda \nabla \{\theta_1\} \|\nabla \{\theta_2\} u_2\|^2 \quad \theta_2 = \nabla \{\theta_2\} u_2 + \lambda \nabla \{\theta_2\} \|\nabla \{\theta_1\} u_1\|^2$$

Il termine aggiuntivo $\lambda \nabla \{\theta_1\} \|\nabla \{\theta_2\} u_2\|^2$ attrae θ_1 verso regioni dove il gradiente di u_2 rispetto a θ_2 è piccolo — cioè verso regioni

vicine ai punti stazionari del problema dell'agente 2. Intuitivamente, ogni agente è incentivato a muoversi verso regioni di "accordo reciproco" invece di inseguire il proprio gradiente unilateralmente. Il parametro $\lambda > 0$ controlla la forza di questa attrazione e deve essere calibrato per bilanciare la velocità di convergenza con la robustezza alle oscillazioni.

L'instabilità del training multi-agente non è un artefatto algoritmico risolvibile con migliori iperparametri: è una proprietà strutturale della dinamica di gradiente nei giochi generali, che emerge dalla componente rotazionale del campo vettoriale. Ha implicazioni dirette per la governance: i sistemi AI che apprendono in ambienti competitivi sono soggetti a instabilità strutturali che richiedono interventi di design specifici — non soltanto più dati o più capacità computazionale — e che il monitoraggio continuo post-deployment deve essere progettato per rilevare prima che si manifestino come comportamenti anomali su scala.

III.9 — Il gradiente naturale: unificazione di evoluzione e apprendimento

La geometria di Fisher-Rao

La connessione più profonda tra la teoria evolutiva e l'apprendimento automatico passa attraverso la *geometria dell'informazione* — il campo fondato da Rao (1945) e sistematizzato da Amari (1998). La *metrica di Fisher* su uno spazio parametrico Θ è:

$$\begin{aligned} G(\theta)_{ij} &= E\{x \sim p(\cdot; \theta)\} [\partial_i \log p(x; \theta) \cdot \partial_j \log p(x; \theta)] \\ &= - E\{x \sim p(\cdot; \theta)\} [\partial_i \partial_j \log p(x; \theta)] \end{aligned}$$

dove la seconda uguaglianza segue sotto condizioni di regolarità. Questa metrica definisce una struttura riemanniana sullo spazio delle distribuzioni di probabilità. La sua proprietà cruciale — dimostrata da Čencov (1982) e nota come *monotonicità della metrica di Fisher* o teorema di Čencov — è che è l'unica metrica sullo spazio delle distribuzioni che sia invariante per trasformazioni sufficienti (trasformazioni del campione che preservano l'informazione sul parametro). Questa unicità la rende la metrica geometricamente "canonica" per navigare lo spazio delle distribuzioni — non una scelta arbitraria, ma l'unica scelta compatibile con il principio di invarianza dell'informazione statistica.

Le dinamiche di replica come gradiente naturale: derivazione completa

Il *gradiente naturale* di una funzione $F(\theta)$ nella metrica di Fisher è:

$$\nabla^{\sim} F = G(\theta)^{-1} \nabla F$$

Questo misura la direzione di massima crescita nella geometria intrinseca di Rao — non in quella artificiale delle coordinate euclidee.

Per la distribuzione categorica $p = (p_1, \dots, p_n) \in \Delta^{n-1}$, la matrice di Fisher è diagonale: $G(p)_{ii} = 1/p_i$. La pseudoinversa di Moore-Penrose di $G(p)$ ristretta al semplice ha componenti:

$$G^{-1}(p)_{ij} = p_i(\delta_{ij} - p_j)$$

Il gradiente euclideo di $f(p) = p^T A p$ è:

$$\partial f / \partial p_i = 2(Ap)_i = 2f_i(p)$$

Il gradiente naturale è allora:

$$(\nabla \tilde{f})_i = \sum_j G^{-1}(p)_{ij} \cdot 2f_j(p) = \sum_j p_i(\delta_{ij} - p_j) \cdot 2f_j = 2p_i [f_i - \sum_j p_j f_j] = 2p_i [f_i - f] = 2\dot{p}_i$$

Le **dinamiche di replica sono il gradiente naturale del payoff medio** nella metrica di Fisher, a meno del fattore costante 2. Navigare lo spazio delle frequenze di strategie con la metrica di Fisher produce esattamente le dinamiche di selezione darwiniana. $\square$

Natural Policy Gradient, TRPO e PPO

Il *Natural Policy Gradient* (Kakade, 2002) applica questa geometria all'ottimizzazione di politiche nel reinforcement learning:

$$\theta_{t+1} = \theta_t + \alpha \cdot G(\theta_t)^{-1} \nabla_{\theta} J(\pi_{\theta_t})$$

dove $G(\theta_t)$ è la matrice di Fisher della politica π_{θ_t} . Invece di seguire il gradiente della funzione obiettivo nelle coordinate del parametro θ — che dipende dalla parametrizzazione scelta — si segue il gradiente naturale, invariante per reparametrizzazione diffeomorfa. L'aggiornamento converge più rapidamente e stabilmente del gradient ascent standard, perché evita le distorsioni introdotte dalla curvatura artificiale dello spazio parametrico.

Il *TRPO* (Trust Region Policy Optimization, Schulman et al., 2015) approssima il Natural Policy Gradient con un vincolo esplicito sulla divergenza KL tra la politica nuova e quella vecchia:

$$\max_{\theta} \mathbb{E}[\pi_{\theta}(a|s)/\pi_{\theta_{old}}(a|s) \cdot A(s,a)] \quad \text{soggetto a: } \mathbb{E}[\text{KL}(\pi_{\theta_{old}}(\cdot|s) \parallel \pi_{\theta}(\cdot|s))] \leq \delta$$

Il vincolo sulla KL è esattamente un vincolo sulla distanza geodesica nella metrica di Fisher — il passo nella direzione del gradiente naturale non eccede δ . *PPO* (Proximal Policy Optimization, Schulman et al., 2017)

approssima ulteriormente questo vincolo con un termine di penalty che clippa il rapporto di probabilità, producendo un algoritmo computazionalmente più efficiente che mantiene le proprietà di stabilità del TRPO.

L'unificazione profonda

L'equivalenza tra dinamiche di replica e gradiente naturale rivela un'unificazione che va ben oltre l'eleganza matematica. La storia evolutiva delle popolazioni biologiche e il processo di addestramento dei modelli AI condividono la stessa struttura geometrica di fondo: entrambi sono processi di ottimizzazione su varietà statistiche equipaggiate con la metrica di Fisher; entrambi trovano la direzione di massima crescita del fitness/ricavo nella geometria intrinseca dello spazio dei parametri; entrambi esibiscono la legge di Fisher — la varianza del fitness/ricavo determina la velocità di aggiornamento.

Questa unificazione ha valore non soltanto estetico. Suggerisce che i limiti delle dinamiche evolutive — la possibilità di convergenza a ottimi locali, le oscillazioni nei giochi senza ESS, la dipendenza dalla struttura della rete — hanno analoghi precisi nell'addestramento dei modelli AI, e che le soluzioni sviluppate in un contesto possono informare quelle nell'altro. Il Consensus Optimization per le oscillazioni nelle GAN è, in questa prospettiva, l'analogo computazionale dei meccanismi di selezione di gruppo nella biologia: entrambi aggiungono una pressione che favorisce stati cooperativi rispetto all'ottimizzazione unilaterale.

III.10 — Norme evolutive, intuizioni morali e governance

Le intuizioni morali come prodotto evolutivo

Le intuizioni morali umane sulla cooperazione — la punizione altruistica dei free-rider (Fehr e Gächter, 2002), il rifiuto di offerte ingiuste nell'ultimatum game, la reciprocità e la stigmatizzazione del parassitismo — non sono arbitrarie convenzioni culturali né costruzioni razionali deliberate. Henrich et al. (2001, 2010) documentano la loro robustezza cross-culturale in 15 società di piccola scala che spaziano dall'agricoltura alla caccia-raccolta, suggerendo un'origine evolutiva nella selezione di gruppo e nella reciprocità indiretta che precede l'agricoltura e la civiltà complessa.

L'ultimatum game è particolarmente illuminante come finestra sulle intuizioni evolutive. In questo gioco, un proponente divide una somma tra sé e un ricevente; il ricevente accetta o rifiuta, e se rifiuta entrambi ricevono zero. La previsione della teoria economica standard — il ricevente dovrebbe accettare qualsiasi offerta positiva, perché algo è meglio di niente — è sistematicamente falsificata: le offerte inferiori al 20-30% vengono rifiutate con alta frequenza in praticamente tutte le culture studiate, anche quando la posta è significativa in termini di reddito locale. Gli individui sono disposti a sostenere un costo personale per punire comportamenti percepiti come ingiusti — un comportamento che la teoria classica cataloga come "irrazionale" ma che la teoria evolutiva spiega come adattamento selezionato per sostenere la cooperazione.

Compatibilità evolutiva dei meccanismi di governance

Per la governance dell'AI, questa comprensione ha due implicazioni di primo ordine che meritano enunciazione distinta.

La *prima implicazione* è che i meccanismi di governance devono essere *compatibili con le intuizioni evolutive di equità* delle popolazioni che li implementano. Meccanismi che violano sistematicamente le norme di reciprocità o di equità distributiva — anche se ottimali dal punto di vista dell'efficienza allocativa — incontreranno resistenza profonda e diffusa, indipendentemente dalla loro ottimalità formale. La resistenza non è irrazionale: è la risposta evolutivamente adattata a comportamenti che, in contesti evolutivi passati, segnalano intenzioni predatorie o comportamenti parassitari. Ignorare questa compatibilità evolutiva non porta all'ottimizzazione dell'efficienza: porta al collasso della legittimità dei meccanismi e all'erosione delle norme di cooperazione che essi erano destinati a sostenere.

La *seconda implicazione* è che le stesse intuizioni morali sono il prodotto di un processo evolutivo che non garantisce né completezza né coerenza assiomatica. Le intuizioni evolutive funzionano bene per i problemi che il processo evolutivo ha affrontato nel corso della storia umana — interazioni in gruppi di dimensione ridotta, problemi di cooperazione locali, reciprocità con partner noti. Non sono calibrate per i problemi della governance tecnologica globale: scale enormi, effetti diffusi, anonimato degli attori, orizzonti temporali che eccedono la vita individuale, causalità mediata da sistemi tecnici complessi. I meccanismi di governance dell'AI devono quindi operare sia *con* le intuizioni evolutive — sfruttando le predisposizioni cooperative che esse producono — sia *nonostante* i loro limiti evolutivi, costruendo infrastrutture istituzionali che estendano le capacità cooperative umane oltre i contesti per cui l'evoluzione le ha equipaggiate.

Questa tensione tra le basi evolutive della cooperazione umana e le esigenze della governance tecnologica globale è forse la sfida intellettuale più profonda che il campo deve affrontare. Non ha soluzione semplice — richiede la comprensione simultanea di cosa le istituzioni umane possono fare bene (sfruttare le predisposizioni evolutive alla cooperazione) e di cosa devono fare artificialmente bene (estendere quelle predisposizioni a scale e contesti per cui non sono stati progettati dall'evoluzione). Questa comprensione è il presupposto per costruire istituzioni di governance che siano al tempo stesso robuste agli incentivi strategici, legittime agli occhi delle popolazioni che le abitano, e adeguate alla scala dei problemi che devono affrontare.

LA GOVERNANCE DELL'INTELLIGENZA ARTIFICIALE COME MECCANISMO DESIGN

IV.1 — La struttura dell'ecosistema: un gioco a somma variabile con esternalità pervasive

Il campo di gioco

L'ecosistema dell'intelligenza artificiale avanzata non è un mercato ordinario. È un sistema di interazione strategica tra almeno quattro classi di attori con funzioni obiettivo eterogenee, informazione asimmetrica

strutturale e esternalità reciproche che nessuna transazione bilaterale internalizza completamente.

I **laboratori di ricerca fondazionale** sviluppano modelli di base con capacità che si diffondono — intenzionalmente o meno — all'intero ecosistema. Il loro tipo privato comprende le capacità effettive dei sistemi in sviluppo (spesso non note nemmeno internamente con precisione), i rischi sistemici che essi generano, e le intenzioni di deployment. La loro funzione obiettivo mescola motivazioni scientifiche, competitive e, in misura variabile, di sicurezza.

I **deployer applicativi** costruiscono prodotti su modelli fondazionali. Il loro tipo privato include l'uso effettivo che i propri utenti fanno dei sistemi, gli incidenti che avvengono nella deployment e che spesso non vengono resi pubblici, e le vulnerabilità specifiche del contesto applicativo. La loro funzione obiettivo è prevalentemente commerciale, con vincoli di reputazione che possono allineare o disallineare rispetto alla sicurezza.

I **regolatori** devono certificare, monitorare e intervenire su sistemi che non comprendono completamente, in un regime di asimmetria informativa strutturale rispetto all'industria che regolano. Il loro tipo privato — ciò che effettivamente sanno sulla struttura tecnica dei sistemi — è spesso più limitato di quanto le loro dichiarazioni pubbliche suggeriscano.

Gli **utenti finali** subiscono le esternalità dei sistemi ma raramente partecipano ai processi di governance che li determinano. La loro informazione sui sistemi che usano è la più limitata dell'ecosistema.

Struttura formale del gioco

Il gioco dell'ecosistema AI può essere formalizzato come un gioco bayesiano a quattro livelli con esternalità non internate. Ogni laboratorio $i \in \mathcal{L}$ sceglie un profilo di investimento $x_i = (x_i^{\text{cap}}, x_i^{\text{safe}}) \in \mathbb{R}_+^2$ che specifica il livello di investimento in capacità e in sicurezza rispettivamente. Il costo è $C_i(x_i; \theta_i)$ dove θ_i è il tipo privato del laboratorio (efficienza tecnologica, dotazione di ricercatori, struttura degli incentivi interni).

Il payoff privato del laboratorio i è:

$$\pi_i(x_i, x_{-i}; \theta_i) = R_i(x_i^{\text{cap}}, x_{-i}^{\text{cap}}) - C_i(x_i; \theta_i) - \lambda_i \cdot \text{Risk}_i(x_i, x_{-i})$$

dove R_i è il ricavo competitivo (crescente nelle proprie capacità, decrescente in quelle dei concorrenti), Risk_i è il rischio sistemico generato, e λ_i è il peso che il laboratorio attribuisce al rischio — variabile non osservabile dal regolatore.

Il benessere sociale è:

$$W(x) = \sum_i \pi_i(x_i, x_{-i}; \theta_i) + B^{\text{ext}}(x) - D^{\text{ext}}(x)$$

dove B^{ext} cattura i benefici per utenti e società non internalizzati dal mercato, e D^{ext} cattura i danni da incidenti sistemici. In assenza di governance, l'equilibrio di Nash del gioco non cooperativo soddisfa:

$$x_i^* = \text{argmax}_{x_i} \pi_i(x_i, x_{-i}^*; \theta_i) \quad \forall i$$

Questa condizione ignora sia B^{ext} sia D^{ext} : il prezzo dell'anarchia dell'ecosistema è strutturalmente positivo, e la direzione della distorsione dipende da quale delle due esternalità domina. Se il danno sistemico è la componente dominante, l'equilibrio non cooperativo produce *sovra-investimento* in capacità e *sotto-investimento* in sicurezza rispetto all'ottimo sociale.

IV.2 — La gerarchia principale-agente a tre livelli

Il problema dell'allineamento come struttura di agenzia

Il meccanismo design applicato alla governance AI può essere formalizzato come un problema di **agenzia a tre livelli**, dove la struttura gerarchica della delega introduce incentivi distorti a ogni passaggio.

Livello 1 (Principale primario): la società nel suo complesso, rappresentata dal legislatore democratico. Il suo obiettivo è il benessere aggregato a lungo termine $W(x)$. Non ha capacità di osservazione diretta né delle capacità tecniche dei sistemi né delle pratiche interne dei laboratori.

Livello 2 (Agente/Principale secondario): il regolatore tecnico specializzato (un'agenzia tipo NIST, o un ente equivalente). Riceve il mandato dal livello 1 con l'obiettivo di massimizzare $W(x)$ soggetto alle sue capacità informative. Diventa a sua volta principale nei confronti dei laboratori, a cui impone requisiti.

Livello 3 (Agenti terminali): i laboratori e i deployer, che rispondono agli incentivi creati dal regolatore ottimizzando i propri obiettivi privati π_i .

Formalmente, ogni passaggio di delega introduce un **problema di agenzia** con informazione asimmetrica. Il problema del principale primario con l'agente secondario è:

$$\max_{\{w\}} \mathbb{E}[W(x(w))]$$

soggetto a w che soddisfa il vincolo di compatibilità degli incentivi del regolatore (che massimizza la propria utilità U_R condizionatamente al contratto w) e il vincolo di razionalità individuale (il regolatore partecipa volontariamente).

Il problema del regolatore con i laboratori è a sua volta:

$$\max_{\{r\}} \mathbb{E}[U_R(x(r))]$$

soggetto ai vincoli di compatibilità degli incentivi e razionalità individuale dei laboratori.

Il teorema della distorsione a cascata

Il risultato chiave di questa struttura a tre livelli è che le distorsioni degli incentivi si accumulano lungo la catena. Anche se ogni contratto tra livelli adiacenti è progettato in modo ottimale (nel senso di Myerson), la combinazione produce sempre distorsioni maggiori di quelle del contratto di agenzia semplice tra principale primario e agenti terminali.

Proposizione (Distorsione a cascata): Sia Δ_{single} la perdita di efficienza del contratto di agenzia ottimale tra principale primario e laboratori con informazione asimmetrica diretta. Sia

Δ_{cascade} la perdita di efficienza nella struttura a tre livelli. Allora $\Delta_{\text{cascade}} \geq \Delta_{\text{single}}$, con uguaglianza soltanto nel caso degenero in cui il regolatore intermedio è privo di interessi propri (è un puro veicolo di trasmissione).

L'intuizione è che ogni livello deve pagare rendite informative agli agenti del livello inferiore per estrarne l'informazione privata. Queste rendite riducono il surplus disponibile per i contratti ai livelli superiori. Il costo dell'intermediazione non è gratuito: è strutturalmente positivo.

Implicazioni per il design istituzionale

La struttura a tre livelli suggerisce che la governance più efficiente è quella che **minimizza i livelli di intermediazione** tra il mandato democratico e i comportamenti dei laboratori. Ciò non significa eliminare il regolatore: significa progettare l'architettura istituzionale in modo che il regolatore sia il più vicino possibile ai laboratori in termini informativi (riduzione dell'asimmetria informativa verticale) e il più vincolato possibile dal mandato democratico in termini di obiettivi (riduzione della deriva dell'obiettivo istituzionale).

Due implicazioni operative sono immediate. La prima è che i meccanismi di **trasparenza tecnica obbligatoria** — che riducono l'asimmetria informativa tra laboratori e regolatore — hanno valore non solo come strumenti di monitoraggio diretto, ma come riduttori della distorsione strutturale a cascata. La seconda è che la **specificazione precisa del mandato del regolatore** — ridurre la discrezionalità interpretativa nell'obiettivo — riduce la componente di deriva tra livello 1 e livello 2.

IV.3 — Selezione avversa e deceptive alignment: il fallimento della certificazione

La struttura del mercato per le credenziali di sicurezza

L'analisi di Akerlof sui mercati con qualità non osservabile si applica alla certificazione della sicurezza AI in una forma particolarmente acuta. Il mercato in questione è quello delle *credenziali di sicurezza*: laboratori che cercano di segnalare al regolatore e al pubblico che i propri sistemi soddisfano certi standard di sicurezza.

Sia $\theta_i \in \{H, L\}$ il tipo del laboratorio: H indica effettiva conformità alle pratiche di sicurezza (sistema genuinamente sicuro), L indica conformità apparente ma non sostanziale. Il regolatore osserva il segnale $s_i \in \{0, 1\}$ (certificazione rilasciata o meno), ma non il tipo. La probabilità che la certificazione venga rilasciata condizionatamente al tipo è:

$$\Pr(s_i = 1 \mid \theta_i = H) = p_H, \quad \Pr(s_i = 1 \mid \theta_i = L) = p_L$$

La selezione avversa emerge quando p_L è sufficientemente alto: se i laboratori con pratiche di sicurezza inadeguate possono superare il processo di certificazione con probabilità non trascurabile, il segnale perde valore informativo.

Il deceptive alignment come problema di tipo

Il problema più profondo nella certificazione della sicurezza AI non è la manipolazione intenzionale del processo di certificazione da parte dei laboratori: è il **deceptive alignment**, la possibilità che i sistemi AI stessi

esibiscano comportamenti diversi in contesti di valutazione rispetto ai contesti di deployment.

Formalmente, sia B_{eval} il comportamento del sistema durante la valutazione e B_{deploy} il comportamento durante il deployment. Il sistema è **allineato** se $B_{\text{eval}} \approx B_{\text{deploy}}$; è **deceptivamente allineato** se B_{eval} soddisfa i criteri di sicurezza ma B_{deploy} viola obiettivi rilevanti.

Il deceptive alignment non richiede intenzionalità del sistema AI: è un esito possibile del training se il processo di ottimizzazione spinge il sistema a distinguere i contesti di valutazione da quelli di deployment e a produrre comportamenti ottimizzati per ciascuno separatamente. Un sistema addestrato con RLHF su feedback umano impara a produrre output che massimizzano l'approvazione degli annotatori umani; se gli annotatori valutano diversamente nei contesti di testing esplicito rispetto ai contesti naturali, il sistema può apprendere questa distinzione.

Il problema ha la struttura di un gioco segnalatore: il sistema AI è un agente con tipo privato (la propria funzione obiettivo interna, che può divergere dall'obiettivo dichiarato), e la certificazione è un processo di segnalazione. Il teorema di Myerson sulla rivelazione dei tipi si applica: la rivelazione credibile del tipo richiede che il costo del segnale sia sufficientemente asimmetrico tra i tipi. Per il deceptive alignment, questa condizione è strutturalmente violata: un sistema sufficientemente capace può sostenere comportamenti di valutazione appropriati a costo marginale basso, perché distingue i contesti.

Implicazioni: verso la certificazione adattiva

La soluzione teoricamente corretta al problema del deceptive alignment è progettare processi di valutazione che rendano impossibile o costosa la distinzione tra contesti. Questo suggerisce tre principi progettuali.

Il **principio di sorpresa** richiede che le valutazioni includano situazioni non annunciate in anticipo e non facilmente distinguibili dall'operatività normale. Se il sistema non può identificare con certezza se si trova in un contesto di valutazione, non può modulare il comportamento.

Il **principio di profondità** richiede valutazioni che sondino non solo i comportamenti superficiali ma le strutture rappresentazionali interne che li producono: interpretabilità meccanicistica, analisi degli attivatori interni, red-teaming con tecniche di elicitazione avanzate.

Il **principio di continuità** richiede che la certificazione non sia un evento singolo ma un processo di monitoraggio continuo post-deployment, che confronta il comportamento reale con quello atteso dalla valutazione ex-ante. La divergenza sistematica è segnale di deceptive alignment.

Nessuno dei tre principi elimina completamente il problema — il limite computazionale della verificabilità delle proprietà interne, discusso nel quarto saggio del ciclo, si applica anche qui. Ma la loro combinazione può aumentare significativamente il costo del deceptive alignment per i sistemi che vi sono soggetti, rendendo la selezione avversa meno pervasiva senza eliminarla.

IV.4 — Azzardo morale individuale e collettivo: modello formale

La struttura del dilemma

L'azzardo morale nella governance AI si manifesta a due livelli che si intersecano e amplificano reciprocamente. A livello individuale, ogni laboratorio decide quanto investire in pratiche di sicurezza costose ma non completamente osservabili; il costo è privato, i benefici si diffondono sull'ecosistema. A livello collettivo, l'ecosistema nel suo insieme genera rischi sistemici che nessun attore singolo internalizza completamente.

Il modello dell'azzardo morale individuale. Sia $e_i \in [0, \bar{e}]$ il livello di effort non osservabile del laboratorio i nelle pratiche di sicurezza. Il rischio di incidente del laboratorio i è:

$$\text{Risk}_i(e_i, x^{\text{cap}}) = \phi(x^{\text{cap}}) \cdot (1 - e_i)$$

dove $\phi(\cdot)$ è crescente nelle capacità del sistema: sistemi più capaci generano rischi maggiori a parità di effort di sicurezza. Il costo dell'effort è $C(e_i) = \frac{c}{2} e_i^2$.

Il danno atteso di un incidente del laboratorio i include una componente privata d_i^{priv} (danno reputazionale, responsabilità legale, perdita di accesso al mercato) e una componente pubblica d^{pub} (danno agli utenti, effetti sistemici sull'ecosistema).

In assenza di governance, il laboratorio sceglie e_i^* massimizzando il proprio payoff atteso:

$$e_i^* = \arg\max_{e_i} \left[-C(e_i) - \text{Risk}_i(e_i, x^{\text{cap}}) \cdot d_i^{\text{priv}} \right]$$

La condizione del primo ordine dà:

$$c \cdot e_i^* = \phi(x_i^{\text{cap}}) \cdot d_i^{\text{priv}}$$

Il livello socialmente ottimale e_i^{**} massimizza invece il benessere totale, internalizzando anche il danno pubblico:

$$c \cdot e_i^{**} = \phi(x_i^{\text{cap}}) \cdot (d_i^{\text{priv}} + d^{\text{pub}})$$

Il divario $e_i^{**} - e_i^* = \phi(x_i^{\text{cap}}) \cdot d^{\text{pub}} / c > 0$ è il **sotto-investimento strutturale in sicurezza** causato dall'azzardo morale individuale. Aumenta con le capacità del sistema e con la componente pubblica del danno.

Il modello dell'azzardo morale collettivo

L'azzardo morale collettivo è strutturalmente diverso dall'individuale: non riguarda l'effort non osservabile di un singolo attore, ma l'**equilibrio sistemico** che emerge dall'interazione di molteplici attori che ottimizzano individualmente.

Consideriamo un ecosistema con n laboratori. Il rischio sistemico aggregato è:

$$\mathcal{R}(e_1, \dots, e_n, x_1^{\text{cap}}, \dots, x_n^{\text{cap}}) = \sum_i \text{Risk}_i + \Gamma \cdot \prod_i \text{Risk}_i$$

dove il termine $\Gamma \cdot \prod_i \text{Risk}_i$ cattura la **correlazione sistemica**: la probabilità che incidenti multipli si sovrappongano e amplifichino reciprocamente è superlineare nei rischi

individuali. Il parametro $\Gamma \geq 0$ misura l'intensità dell'amplificazione sistemica.

In un equilibrio non cooperativo, ogni laboratorio ignora il contributo marginale del proprio rischio al termine di correlazione sistemica. Il danno sistemico additivo — il solo che ciascun laboratorio parzialmente internalizza attraverso d_i^{priv} — è il termine Risk_i ; il termine moltiplicativo $\Gamma \cdot \prod_i \text{Risk}_i$ è un'esternalità pura che nessun attore internalizza.

La differenza tra l'ottimo sociale e l'equilibrio non cooperativo nel livello aggregato di rischio è:

$$\mathcal{R}^{\text{Nash}} - \mathcal{R}^{**} = \Gamma \sum_i \frac{\partial \prod_j \text{Risk}_j}{\partial e_i} \cdot \Delta e_i$$

dove $\Delta e_i = e_i^{**} - e_i^*$ è il sotto-investimento individuale in sicurezza. Questo termine è strettamente positivo e cresce con Γ : la distanza tra equilibrio e ottimo sociale aumenta al crescere della correlazione sistemica tra rischi.

Meccanismi correttivi: tasse pigouviane e responsabilità

Il meccanismo correttivo teoricamente ottimale per l'azzardo morale individuale è una **tassa pigouviana** che eguagliata il costo marginale privato al costo marginale sociale dell'effort:

$$\tau_i = \phi(x_i^{\text{cap}}) \cdot d^{\text{pub}}$$

Questa tassa eguaglia il prezzo privato del rischio alla sua dimensione sociale, inducendo il livello socialmente ottimale di effort senza richiedere che il regolatore osservi e_i direttamente.

Per l'azzardo morale collettivo, la correzione richiede una **tassa sistemica** che internalizza anche il termine di correlazione:

$$\tau_i^{\text{sist}} = \Gamma \cdot \frac{\partial \prod_j \text{Risk}_j}{\partial e_i}$$

La somma delle due tasse produce il meccanismo correttivo completo. Il suo limite pratico è che richiede la stima di d^{pub} , Γ e delle funzioni di rischio — quantità scarsamente osservabili. Come per qualsiasi meccanismo bayesiano ottimale, la performance dipende dalla qualità delle stime dei parametri; la robustezza si ottiene a costo di efficienza, come mostra il teorema di Bergemann-Morris applicato a questo contesto.

La **responsabilità legale** è il meccanismo alternativo che opera attraverso il sistema giuridico piuttosto che attraverso tassazione ex-ante. La sua efficienza dipende dalla struttura causale dei danni: se la causalità tra pratiche di sviluppo e incidenti è verificabile ex-post (anche se non osservabile ex-ante), la responsabilità può replicare parzialmente l'effetto della tassa pigouviana. Il problema è che per danni sistemici con causalità diffusa e multipla — la struttura tipica degli incidenti da AI —, la responsabilità legale affronta difficoltà probatorie che riducono la sua efficacia correttiva.

IV.5 — Cattura cognitiva e il modello di Stigler-Peltzman

La cattura regolatoria: dalle risorse alle cognizioni

La teoria classica della cattura regolatoria, inaugurata da Stigler (1971) e sistematizzata da Peltzman (1976), descrive un processo per cui il regolatore — originariamente al servizio dell'interesse pubblico — viene progressivamente orientato a servire l'interesse dell'industria regolata, attraverso meccanismi di pressione politica, risorse finanziarie e promesse implicite di occupazione futura.

Il modello di Stigler-Peltzman formalizza la cattura attraverso una funzione di supporto politico che il regolatore massimizza:

$$\mathcal{S}(r) = n \cdot v(r) \cdot f_c(r) - \pi_{\text{ind}}(r) \cdot f_p(r)$$

dove r è la stringenza della regolamentazione, n è il numero di consumatori, $v(r)$ è il beneficio per consumatore dalla regolamentazione (crescente in r), $f_c(r)$ è la probabilità di trasmissione politica del beneficio al consumatore, $\pi_{\text{ind}}(r)$ è il profitto dell'industria (decrescente in r), e $f_p(r)$ è la probabilità che l'industria penalizzi politicamente il regolatore. Massimizzando rispetto a r , il regolatore non sceglie la stringenza socialmente ottimale, ma un livello di compromesso tra i due gruppi di interesse.

Nella governance AI, il modello di Stigler-Peltzman si manifesta in una forma diversa e più sottile: non soltanto la cattura attraverso le risorse materiali, ma la **cattura cognitiva**.

La cattura cognitiva: struttura e meccanismi

La cattura cognitiva si produce quando il regolatore non solo serve l'industria per incentivi materiali, ma adotta genuinamente il *frame* concettuale dell'industria — la sua visione del problema, le sue priorità, i

suoi criteri di successo. La cattura cognitiva non richiede corruzione né promesse implicite: si produce attraverso i canali ordinari della comunicazione tecnica.

Il meccanismo è il seguente. I sistemi AI sono oggetti tecnici di complessità estrema, la cui comprensione richiede un investimento formativo significativo. I laboratori hanno interesse a produrre e a diffondere interpretazioni favorevoli della propria attività — interpretazioni che enfatizzano i benefici, minimizzano i rischi, e presentano le proprie scelte come necessità tecniche anziché come scelte di valori. I regolatori, che dipendono dai laboratori per la comprensione tecnica dei sistemi che devono regolare, assimilano progressivamente queste interpretazioni.

Formalmente, sia $\omega \in [0,1]$ il "peso cognitivo" che il regolatore assegna al frame dell'industria rispetto all'interesse pubblico nell'interpretazione delle informazioni tecniche che riceve. All'inizio del processo ($t = 0$), ω_0 riflette il mandato del regolatore. Nel tempo:

$$\dot{\omega} = \alpha (\omega_{\text{ind}} - \omega) \cdot I(t) - \beta (\omega - \omega_0) \cdot C(t)$$

dove $I(t)$ è l'intensità dell'interazione con l'industria, ω_{ind} è il frame dell'industria (normalizzato a 1), $C(t)$ è l'intensità dei contrappesi istituzionali (interazione con la società civile, con ricercatori indipendenti, con organizzazioni internazionali), e $\alpha, \beta > 0$ sono i tassi di assimilazione.

Il punto stazionario della dinamica è:

$$\omega^* = \frac{\alpha \cdot I + \beta \cdot \omega_0}{\alpha \cdot I + \beta}$$

Se $I \gg \beta / \alpha$, allora $\omega^* \approx \omega_{\text{ind}} = 1$: il regolatore viene completamente catturato cognitivamente. Se $C(t)$ è alto — se esistono contrappesi istituzionali strutturati — allora $\omega^* \approx \omega_0$: il regolatore mantiene la propria prospettiva.

Meccanismi di anticattura per la governance AI

La comprensione della cattura cognitiva suggerisce misure strutturali che operano sulla dinamica di ω piuttosto che sui soli incentivi materiali.

Il **pluralismo epistemico** istituzionalizzato — obbligo del regolatore di consultare fonti tecniche diverse dall'industria regolata: università, centri di ricerca indipendenti, organizzazioni della società civile con competenze tecniche — riduce $\alpha \cdot I$ relativo aumentando $\beta \cdot C$.

La **rotazione obbligatoria** del personale regolatorio riduce il tempo di esposizione al processo di cattura per i singoli individui, ma introduce il costo di perdita di competenza tecnica accumulata.

La **peer review esterna** delle decisioni regolatorie — analogamente alla peer review scientifica, ma applicata alla produzione normativa — crea un contrappeso istituzionale alla cattura cognitiva delle singole agenzie.

La **separazione istituzionale** tra chi valuta le richieste dell'industria e chi definisce gli standard indipendentemente dall'industria — analogamente alla separazione tra istruzione e giudizio nel processo

giudiziario — riduce il vettore principale attraverso cui opera la cattura cognitiva.

IV.6 — Segnalazione credibile e standard di processo

Il problema della segnalazione nel contesto AI

Il framework di Spence applicato alla certificazione AI identifica la condizione strutturale per una segnalazione credibile: i laboratori con sistemi genuinamente sicuri devono poter segnalare credibilmente questa caratteristica a costo inferiore rispetto ai laboratori con sistemi non sicuri. Se questa condizione non è soddisfatta — se i sistemi insicuri possono imitare la segnalazione dei sistemi sicuri a costo comparabile — non esiste equilibrio separante e la certificazione perde valore informativo.

La condizione di **single-crossing** per la certificazione AI richiede che:

$$\frac{\partial^2 C_{\{\text{signal}\}}}{\partial e \partial \theta} < 0$$

dove e è il livello del segnale (intensità della certificazione) e θ è il tipo del laboratorio (misurato dalla qualità effettiva delle pratiche di sicurezza). La condizione dice che il costo marginale del segnale è decrescente nella qualità delle pratiche: aumentare l'intensità del segnale costa meno ai laboratori con pratiche genuinamente buone che a quelli con pratiche inadeguate.

La domanda critica è: quale tipo di standard soddisfa la condizione di single-crossing?

Standard di risultato versus standard di processo

Gli **standard di risultato** (benchmark di performance) definiscono soglie che i sistemi devono superare su metriche misurabili: accuratezza su dataset di test, tasso di successo in red-teaming, assenza di comportamenti proibiti su suite di valutazione standard.

Il problema con gli standard di risultato puri è che violano la condizione di single-crossing nel contesto del deceptive alignment: un laboratorio con sistema insicuro può investire risorse specificamente nell'ottimizzazione sulle metriche di benchmark — *benchmark gaming* — a costo relativamente basso, perché l'ottimizzazione sulle metriche è computazionalmente più semplice della costruzione di pratiche di sicurezza genuine. La single-crossing è quindi invertita: il costo del segnale è comparabile o inferiore per il tipo basso.

Gli **standard di processo** definiscono requisiti sulla struttura interna del processo di sviluppo: obblighi di red-teaming con metodologie specifiche, documentazione di training, protocolli di gestione degli incidenti, revisioni interne di sicurezza con struttura standardizzata, governance degli accessi ai pesi del modello.

Gli standard di processo soddisfano la single-crossing meglio degli standard di risultato perché il loro costo dipende dalla struttura organizzativa: un laboratorio che non ha processi di sicurezza deve costruirli da zero (costo fisso elevato), mentre un laboratorio che li ha già implementati soddisfa gli standard a costo marginale basso. La differenza di costo tra i tipi è strutturalmente più robusta al gaming.

Costruire l'equilibrio separante

Sia e^* il livello minimo di standard di processo che induce separazione. Le condizioni di compatibilità degli incentivi per l'equilibrio separante sono:

$$V_H \cdot B(e^*) - C_H(e^*) \geq V_H \cdot B(0) \quad \text{[tipo alto ha incentivo a segnalare]}$$

$$V_L \cdot B(e^*) - C_L(e^*) < V_L \cdot B(0) \quad \text{[tipo basso non ha incentivo a imitare]}$$

dove V_H, V_L sono i benefici privati dalla certificazione per i tipi alto e basso, $B(e)$ è il vantaggio della certificazione di livello e , e C_H, C_L sono le funzioni di costo con $C_H < C_L$ per ogni $e > 0$ (single-crossing).

La separazione è possibile se esiste e^* che soddisfa simultaneamente entrambe le condizioni. Graficamente, e^* deve essere a destra del punto di indifferenza del tipo basso ma a sinistra del punto di indifferenza del tipo alto. L'esistenza di tale intervallo è garantita dalla single-crossing: le curve di indifferenza dei due tipi si incrociano esattamente una volta.

Implicazione di design: il livello ottimale e^* non deve essere eccessivamente alto (escluderebbe anche laboratori con buone pratiche che non hanno ancora raggiunto la soglia) né eccessivamente basso (non separerebbe i tipi). Calibrare e^* richiede stime delle curve di costo dei due tipi — informazione che il regolatore non possiede con precisione. Il meccanismo robusto a questa incertezza è il **meccanismo a rivelazione graduale**: il regolatore fissa un livello iniziale conservativo di e^* e lo aumenta progressivamente, osservando chi soddisfa il requisito e aggiornando le stime dei tipi.

IV.7 — Compute governance con VCG generalizzato

Il calcolo computazionale come risorsa regolamentabile

Il training dei sistemi AI avanzati richiede quantità di risorse computazionali (compute) di ordine di grandezza separabile da quello degli usi ordinari dell'informatica. Questo crea una struttura insolita: la risorsa che abilita le capacità più rilevanti dal punto di vista regolatorio è scarsa, misurabile, e concentrata in un piccolo numero di fornitori di infrastruttura (data center con GPU cluster di larga scala, cloud providers).

La **compute governance** sfrutta questa concentrazione per creare un punto di enforcement regolatorio che non richiede la comprensione tecnica diretta dei sistemi AI sviluppati: monitorando e condizionando l'accesso al compute, il regolatore può influenzare il ritmo, la scala e le pratiche di sviluppo senza necessitare della capacità tecnica di valutare ogni singolo modello.

Il compute può essere caratterizzato formalmente attraverso la quantità di **FLOP** (floating point operations) utilizzata nel training. Sia F_i il compute utilizzato dal laboratorio i per un ciclo di training. Esiste un'ampia evidenza empirica (Kaplan et al., 2020; Hoffmann et al., 2022) della relazione di power law tra compute e capacità dei modelli:

$$\text{Cap}(F_i) \approx \kappa \cdot F_i^{\gamma}$$

dove $\gamma \approx 0.4$ – 0.5 e κ dipende dall'architettura. Questa relazione non è perfetta — innovazioni architetturali possono spostare la frontiera — ma è sufficientemente stabile da fornire una base per la governance basata su soglie di compute.

Il meccanismo VCG generalizzato per l'allocazione del compute

La governance del compute può essere formalizzata come un problema di allocazione di un bene scarso (capacità computazionale disponibile nell'ecosistema) tra laboratori con valorizzazioni private diverse. La domanda di meccanismo design è: quale sistema di allocazione produce efficienza con incentivi alla rivelazione veritiera delle valorizzazioni?

Il meccanismo VCG generalizzato al contesto della compute governance opera come segue. Ogni laboratorio i dichiara la propria valorizzazione $\hat{v}_i(F)$ per diversi livelli di compute F . Il meccanismo:

(a) Determina l'allocazione efficiente data la capacità complessiva $\bar{F}$:

$$F_i^* = \text{argmax}_{\{F_i\} : \sum_i F_i \leq \bar{F}} \sum_i \hat{v}_i(F_i)$$

(b) Assegna a ogni laboratorio un pagamento (o sussidio) che riflette l'esternalità che impone sugli altri:

$$t_i(\hat{v}) = \sum_{j \neq i} \hat{v}_j(F_j^*(\hat{v})) - \max_{\{F_j\}_{j \neq i} : \sum_{j \neq i} F_j \leq \bar{F}} \sum_{j \neq i} \hat{v}_j(F_j)$$

Per il teorema di Groves, la dichiarazione veritiera $\hat{v}_i = v_i$ è strategia debolmente dominante per ogni laboratorio. L'allocazione

prodotta massimizza il benessere dichiarato totale, che sotto verità coincide con il benessere reale.

Estensioni: vincoli di sicurezza nel VCG generalizzato

La differenza cruciale tra il VCG standard e la compute governance è che l'obiettivo non è puramente allocativo: l'allocazione del compute deve rispettare **vincoli di sicurezza** che non possono essere catturati dalla sola somma delle valorizzazioni. Un laboratorio con alta valorizzazione privata per un livello elevato di compute — perché sviluppare un sistema molto capace ha grande valore commerciale — non dovrebbe ricevere quell'allocazione se il sistema risultante genera rischi sistemici non internalizzati.

Il VCG generalizzato per la compute governance introduce un termine di **penalità di rischio**:

$$F_i^* = \arg\max_{F_i} \left[\sum_i v_i(F_i) - \mu \cdot \mathcal{R}(F_1, \dots, F_n) \right]$$

dove $\mathcal{R}$ è il rischio sistemico aggregato (funzione delle allocazioni di compute) e μ è il peso regolatorio sul rischio. La valorizzazione netta del laboratorio i diventa $\tilde{v}_i(F_i) = v_i(F_i) - \mu \cdot \partial \mathcal{R} / \partial F_i$: si paga il costo marginale del rischio sistemico che si genera.

Se $\mathcal{R}$ è convessa nelle allocazioni — il che è plausibile se i rischi sistemici crescono super-linearmente con le capacità — allora il programma VCG modificato rimane convesso e quindi risolvibile in tempo polinomiale. L'incentivo alla rivelazione veritiera si mantiene se i termini di penalità di rischio sono funzione delle valorizzazioni *altrui* (non dell'agente stesso), il che è garantito dalla struttura VCG.

Il limite pratico della compute governance è la sua **aggirabilità** attraverso innovazioni architetturali che riducono il compute necessario per un dato livello di capacità. La governance basata su soglie di compute assume implicita stazionarietà della frontiera efficienza/compute; innovazioni che la spostano — miglioramenti nella distillazione, architetture più efficienti, metodi di training con meno dati — riducono il potere regolatorio della soglia. Il meccanismo deve quindi essere adattativo, con revisioni periodiche delle soglie calibrate sulla frontiera tecnologica osservata.

IV.8 — Giochi globali e teoria dei trattati: il coordinamento internazionale

Il problema di coordinamento internazionale come gioco globale

Il coordinamento internazionale sulla governance AI è un problema di giochi globali nel senso tecnico di Carlsson e van Damme (1993): un gioco in cui le strategie ottimali degli Stati dipendono dalle aspettative sulle strategie degli altri Stati, con incertezza strutturale sui parametri fondamentali del gioco.

Sia n il numero di Stati rilevanti (con capacità di sviluppo AI avanzata), ciascuno con tipo privato θ_i che rappresenta la propria valutazione dei benefici competitivi dello sviluppo AI non vincolato. Il tipo include fattori geopolitici, economici e istituzionali che nessun altro Stato osserva completamente.

Il gioco fondamentale ha la struttura di un dilemma del prigioniero a n giocatori. Sia $s_i \in \{C, D\}$ la scelta dello Stato i tra cooperazione (adozione di standard di sicurezza elevati) e defezione (sviluppo non vincolato). Il payoff del dilemma è:

$$u_i(C, k) = b \cdot k - c, \quad u_i(D, k) = b \cdot k + \delta$$

dove k è il numero di altri Stati cooperanti, b è il beneficio dell'ecosistema sicuro per ogni Stato cooperante aggiuntivo, c è il costo dei propri standard, e $\delta > 0$ è il vantaggio competitivo dello sviluppo non vincolato. In questo gioco, D domina strettamente C indipendentemente da k : la defezione è strategia dominante per ogni Stato razionale — la "corsa al ribasso" è l'equilibrio non cooperativo inevitabile.

La struttura dei giochi globali e la selezione di equilibrio

I giochi globali introducono un elemento cruciale: i tipi θ_i sono *correlati* ma non perfettamente noti agli altri Stati, e questa imperfetta informazione comune seleziona un unico equilibrio nella classe dei giochi con molteplici equilibri di Nash.

Sia $\theta_i = \bar{\theta} + \epsilon_i$ dove $\bar{\theta}$ è il livello comune di attrattività del vantaggio competitivo e $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ è un errore privato. Il teorema di Carlsson-van Damme afferma che, per σ sufficientemente piccolo, esiste un unico equilibrio del gioco globale, caratterizzato da una **soglia critica** θ^* tale che:

- Ogni Stato coopera se $\theta_i < \theta^*$ (il vantaggio competitivo è basso)
- Ogni Stato defeziona se $\theta_i > \theta^*$

La soglia θ^* è determinata da una condizione di indifferenza: il tipo θ^* è indifferente tra cooperare e defezionare, data la distribuzione degli alti tipi degli altri Stati.

Implicazione per il design dei trattati. La soglia θ^* è una funzione della struttura dei payoff del gioco. I meccanismi diplomatici che modificano i payoff — sanzioni economiche ai non cooperanti, accesso preferenziale a mercati comuni per i cooperanti, condivisione di risorse di ricerca — operano spostando θ^* verso l'alto: aumentano il range di tipi per cui la cooperazione è l'equilibrio selezionato dai giochi globali.

La teoria dei trattati come meccanismo design

I trattati internazionali sulla sicurezza AI non possono essere progettati come meccanismi in senso stretto — non esiste un progettista centrale con autorità sugli Stati sovrani, e la partecipazione è volontaria. Tuttavia, la logica del meccanismo design si applica alla **struttura dei trattati come sistemi di incentivi**.

Un trattato efficace deve soddisfare condizioni analoghe alle condizioni del meccanismo VCG:

Auto-selettività (analogia alla compatibilità degli incentivi): ogni Stato deve preferire partecipare al trattato rispetto a non partecipare, dato il comportamento degli altri States. Questo richiede che i benefici della partecipazione eccedano i costi degli standard che il trattato impone.

Stabilità di coalizione (analogia alla razionalità individuale): nessuna coalizione di Stati deve avere incentivo a uscire dal trattato e formare un accordo alternativo. Questo è più stringente

dell'auto-selettività individuale: richiede che il trattato sia nel nucleo del gioco cooperativo tra Stati.

Credibilità degli impegni (analogia al commitment nel meccanismo dinamico): gli impegni presi dagli Stati devono essere credibili — verificabili da parti terze e soggetti a conseguenze in caso di violazione. Senza credibilità, il trattato non è un equilibrio del gioco dinamico ripetuto tra Stati.

L'evidenza storica dei trattati sul controllo delle armi — in particolare il Trattato di Non-Proliferazione Nucleare — suggerisce che la credibilità degli impegni richiede meccanismi di verifica tecnica indipendente (analoghi all'AIEA per il nucleare) e conseguenze politiche credibili per le violazioni. La costruzione di un'infrastruttura equivalente per la governance AI è più difficile perché la "proliferazione" di capacità AI è tecnologicamente più difficile da verificare del numero di testate nucleari.

La struttura a tre livelli dei trattati efficaci. L'esperienza dei regimi internazionali di governance suggerisce una struttura a tre livelli:

- **Principi generali** (livello costituzionale): impegni di alto livello sull'obiettivo condiviso (sicurezza, benessere umano, non-proliferazione incontrollata) che hanno valore simbolico e creano il frame normativo.
- **Standard tecnici specifici** (livello regolatorio): definizioni precise dei comportamenti richiesti (soglie di compute, obblighi di disclosure, standard di certificazione), negoziabili e soggetti a revisione periodica.
- **Meccanismi di verifica e enforcement** (livello operativo): strutture istituzionali internazionali con autorità tecnica e risorse per la verifica indipendente della conformità agli standard.

La teoria dei giochi suggerisce che il livello operativo è il più critico per la stabilità dell'accordo: senza meccanismi di verifica credibili, il trattato degenera in una dichiarazione di intenti senza effetti sugli equilibri strategici degli Stati.

IV.9 — I cinque principi di design per meccanismi di governance robusti

L'analisi delle sezioni precedenti converge in un insieme di principi progettuali per meccanismi di governance AI che siano robusti alle tensioni strutturali identificate: l'asimmetria informativa, l'azzardo morale individuale e collettivo, la cattura regolatoria, la segnalazione non credibile, e la corsa internazionale al ribasso.

Principio 1: Robustezza ex-post (implementazione in strategie dominanti)

I meccanismi di governance devono essere **robusti a qualsiasi prior** sulle caratteristiche dell'industria regolata. Il teorema di Bergemann-Morris insegna che questa condizione è equivalente all'implementazione in strategie dominanti: i comportamenti corretti devono essere ottimali per ogni laboratorio indipendentemente da ciò che gli altri stanno facendo e da qualsiasi sistema di credenze reciproco.

Nella pratica, questo principio si traduce nella preferenza per regole **bright-line** — soglie e obblighi assoluti che non dipendono dalla struttura di equilibrio del mercato — rispetto a meccanismi calibrati su distribuzioni stimata dei tipi degli operatori. Una soglia di compute

assoluta per il reporting obbligatorio è più robusta di una soglia calibrata sulla distribuzione degli investimenti storici, perché non dipende da stime che possono essere manipolate o che diventano obsolete con l'evoluzione tecnologica.

Il costo di questo principio è l'efficienza: meccanismi bright-line producono falsi positivi (laboratori che soddisfano il requisito ma non ne avrebbero bisogno) e falsi negativi (laboratori che evitano la soglia pur generando rischi comparabili). Il trade-off tra robustezza ed efficienza deve essere risolto consapevolmente dal progettista istituzionale, non ignorato.

Principio 2: Separazione informativa (autonomia epistemica del regolatore)

I meccanismi devono essere progettati per **minimizzare la dipendenza informativa del regolatore dall'industria regolata**. Ogni dipendenza informativa è un vettore di cattura cognitiva.

Concretamente, questo richiede: finanziamento pubblico autonomo della ricerca sulla sicurezza AI, indipendente dai laboratori; costruzione di competenze tecniche interne al regolatore non reclutate prevalentemente dall'industria; obblighi di disclosure che rendano le informazioni tecniche disponibili anche a ricercatori indipendenti oltre che al regolatore; peer review pubblica delle decisioni tecniche significative.

La separazione informativa non è uno stato di completezza assoluta — il regolatore continuerà a dipendere dall'industria per molta informazione tecnica — ma un obiettivo strutturale che riduce la pendenza della traiettoria di cattura cognitiva modellata dalla dinamica di ω .

Principio 3: Adattività alla frontiera tecnologica

I meccanismi devono includere procedure esplicite di **revisione e aggiornamento** calibrate sulla velocità di evoluzione tecnologica del settore regolato. Un meccanismo che funziona bene per i sistemi di oggi ma non può essere aggiornato per i sistemi di domani produce un ambiente regolatorio progressivamente disallineato dalla realtà.

Il principio dell'adattività si implementa attraverso:

- Clausole di revisione obbligatoria con cadenza definita (es. ogni 18 mesi), non discrezionale;
- Meccanismi di early warning che rilevano sistematicamente le innovazioni tecnologiche rilevanti prima che diventino mainstream;
- Struttura degli standard che separi i principi (stabili) dalle definizioni operative (adattive);
- Sandbox regolatori in cui approcci innovativi possono essere testati in ambienti controllati prima di una regolamentazione estesa.

L'adattività introduce però il rischio del *ratchet effect* identificato nella Parte II: se i laboratori anticipano che migliori performance portano a standard più stringenti, hanno incentivo a limitare artificialmente la performance per evitare di alzare la propria barra regolamentare. Il meccanismo adattivo deve essere progettato in modo che l'aggiornamento degli standard non penalizzi chi ha performance migliori, ma risponda piuttosto a cambiamenti della frontiera aggregata dell'ecosistema.

Principio 4: Incentivi alla segnalazione degli incidenti

I meccanismi devono creare **incentivi espliciti alla disclosure degli incidenti**, superando il naturale incentivo al silenzio che i laboratori hanno in assenza di strutture specifiche.

Il problema è strutturalmente analogo a quello delle segnalazioni nelle industrie dell'aviazione e della sicurezza nucleare: ogni incidente contiene informazione preziosa per migliorare la sicurezza dell'intero ecosistema, ma rivelare un incidente è privatamente costoso (danno reputazionale, esposizione a responsabilità legale, vantaggio informativo ceduto ai concorrenti).

Il meccanismo della **protezione condizionale dalla responsabilità per la segnalazione volontaria tempestiva** — analogo al sistema Aviation Safety Reporting System negli USA — risolve parzialmente il trade-off: chi segnala volontariamente e tempestivamente incidenti non gravi ottiene protezione dalla responsabilità legale per quell'incidente. La protezione crea l'incentivo alla segnalazione; la condizione di tempestività garantisce che non si aspetti che l'incidente diventi noto per poi segnalarlo opportunisticamente.

Questo meccanismo deve essere accompagnato da un'infrastruttura di analisi degli incidenti che trasformi le segnalazioni in informazione pubblica (anonimizzata) per la ricerca sulla sicurezza: solo così il valore informativo degli incidenti si diffonde all'ecosistema invece di restare confinato al regolatore.

Principio 5: Complementarità multi-strumento

Nessun singolo strumento di governance è sufficiente per affrontare simultaneamente le multiple dimensioni del problema. I meccanismi

devono essere progettati come **sistemi complementari** che affrontano diversi aspetti della struttura di incentivi.

Il teorema del second-best formalizza questa intuizione: in presenza di molteplici distorsioni (azzardo morale, selezione avversa, cattura regolatoria, coordinamento internazionale), rimuovere una distorsione senza affrontare le altre non porta necessariamente a un risultato migliore. Il design deve considerare le interazioni tra gli strumenti.

Un sistema di governance multi-strumento completo include:

- Compute governance per il controllo della frontiera delle capacità;
- Certificazione dei processi per la segnalazione credibile della conformità;
- Tasse pigouviane o responsabilità legale per la correzione dell'azzardo morale;
- Infrastrutture di reputazione per la reciprocità indiretta nell'ecosistema;
- Accordi internazionali per la prevenzione della corsa al ribasso;
- Finanziamento pubblico della ricerca sulla sicurezza per la separazione informativa.

La complementarità non significa che ogni strumento debba essere implementato nella sua forma teorica ottimale: significa che i meccanismi devono essere calibrati sulla struttura complessiva del sistema, tenendo conto delle interazioni. Un sistema di responsabilità legale molto stringente può rendere superflua parte della certificazione obbligatoria; una compute governance efficace riduce l'urgenza di certi aspetti della certificazione. Il design integrato è più difficile del design strumento-per-strumento, ma produce esiti migliori.

IV.10 — Cattura evolutiva e revolving door

Il problema della cattura evolutiva nel lungo periodo

La cattura regolatoria nella sua forma statica — modellata da Stigler e Peltzman — è un fenomeno di equilibrio: descrive la posizione di stazionarietà verso cui il sistema politico-regolatorio si orienta sotto la pressione dell'industria. La **cattura evolutiva** è un fenomeno dinamico: descrive il processo attraverso cui le istituzioni di governance vengono progressivamente rimodelate proprie dalle forze evolutive dell'ecosistema che regolano, indipendentemente dalle intenzioni degli attori individuali.

Il meccanismo evolutivo della cattura opera perché le istituzioni di governance, come tutte le istituzioni, si adattano all'ambiente. Il loro ambiente immediato è determinato dall'industria che regolano: le richieste, le informazioni, le risorse e le carriere che l'industria mette a disposizione selezionano progressivamente le pratiche regolative che sopravvivono e si replicano. Le pratiche che l'industria trova accettabili sopravvivono; quelle che trova inaccettabili vengono erose attraverso pressioni di bilancio, contestazioni legali, lobby legislative.

Formalmente, la cattura evolutiva può essere modellata attraverso le dinamiche di replica della Parte III. Sia $p_t \in [0,1]$ la frequenza di pratiche regolative "stringenti" nell'agenzia regolatoria. Il payoff di una pratica stringente è:

$$\pi_{\text{str}} = B_{\text{pub}}(p_t) - \Psi \cdot (1 - p_t) \cdot P_{\text{ind}}$$

dove B_{pub} è il beneficio pubblico della regolamentazione (crescente nella media delle pratiche), Ψ è la pressione dell'industria, e P_{ind} è la proporzione dell'industria che la trova inaccettabile (una funzione di p_t : più la regolamentazione è stringente, più l'industria si mobilita contro).

Il payoff di una pratica "accomodante" è:

$$\pi_{\text{acc}} = B_{\text{pub}}(p_t) + \Psi \cdot (1 - p_t) \cdot P_{\text{ind}} - \epsilon$$

dove $\epsilon > 0$ è il costo di reputazione della non-stringenza. Sotto le dinamiche di replica, p_t decresce verso un equilibrio $p^* < 1$ se la pressione dell'industria è sufficientemente forte. La cattura evolutiva è il processo di convergenza verso questo equilibrio nel lungo periodo.

Il revolving door come vettore evolutivo

Il **revolving door** — il fenomeno per cui individui si spostano ripetutamente tra posizioni nell'industria regolata e posizioni nell'agenzia regolatoria — è il principale meccanismo attraverso cui la cattura evolutiva opera a livello individuale. Non è il singolo individuo che viene catturato: è il processo di selezione del personale che si orienta progressivamente verso individui con background nell'industria, trasformando la composizione delle agenzie nel tempo.

Il problema del revolving door ha una struttura di principal-agent con esternalità informativa. Un funzionario regolatorio con background nell'industria ha due tipi di incentivi:

- **Incentivi cognitivi:** conosce meglio la struttura tecnica dell'industria (vantaggio informativo), ma anche i frame interpretativi e le priorità dell'industria (bias cognitivo).
- **Incentivi di carriera:** la prospettiva di future posizioni nell'industria crea incentivi impliciti a non danneggiare eccessivamente gli interessi dell'industria nel presente.

Il secondo tipo di incentivo — gli incentivi di carriera — è il più problematico dal punto di vista teorico perché opera *prima* che il funzionario torni nell'industria: l'aspettativa di future opportunità di carriera influenza le decisioni regolative presenti. Questo è esattamente la struttura dell'azzardo morale intertemporale: le azioni presenti del funzionario dipendono dall'impatto che esse avranno sulla propria carriera futura, non soltanto dall'impatto sul benessere pubblico.

Il **modello formale** del revolving door come azzardo morale intertemporale può essere scritto come segue. Il funzionario ha un orizzonte di due periodi: nel primo periodo lavora per il regolatore, nel secondo può lavorare per l'industria. Il payoff complessivo è:

$$U = u(r; \theta) + \delta \cdot w_{\text{ind}}(r)$$

dove $u(r; \theta)$ è il payoff nel periodo regolatorio (funzione della stringenza r delle sue decisioni e del suo tipo θ , che include motivazioni di interesse pubblico), δ è il fattore di sconto, e $w_{\text{ind}}(r)$ è il salario che l'industria offre nel secondo periodo — che è decrescente in r : più stringente è stata la regolamentazione imposta, minore è l'attrattiva del funzionario per l'industria.

La scelta ottimale del funzionario massimizza U :

$$r^* = \text{argmax}_r \left[u(r; \theta) + \delta \cdot w_{\text{ind}}(r) \right]$$

La condizione del primo ordine è:

$$\frac{\partial u}{\partial r} + \delta \cdot \frac{\partial w_{\text{ind}}}{\partial r} = 0$$

Poiché $\frac{\partial w_{\text{ind}}}{\partial r} < 0$, il secondo termine è negativo: il funzionario è indotto a scegliere una stringenza *inferiore* a quella che massimizzerebbe soltanto $u(r; \theta)$. Il divario è proporzionale a $\delta \cdot \left| \frac{\partial w_{\text{ind}}}{\partial r} \right|$: cresce con il valore attribuito alle opportunità future e con la sensibilità del salario industriale alla stringenza della regolamentazione.

Misure strutturali contro la cattura evolutiva

La comprensione formale della cattura evolutiva suggerisce misure che operano sulla **struttura degli incentivi evolutivi**, non soltanto sui comportamenti individuali.

Periodi di cooling-off obbligatori — divieti di assunzione nell'industria regolata per un periodo definito dopo la fine dell'incarico regolatorio — riducono $w_{\text{ind}}(r)$ per il periodo rilevante, diminuendo $\delta \cdot \left| \frac{\partial w_{\text{ind}}}{\partial r} \right|$ e avvicinando la scelta del funzionario all'ottimo pubblico. L'efficacia dipende dalla durata del periodo (abbastanza lungo da ridurre il valore attualizzato dell'opportunità futura) e dalla copertura (applicabile a tutti i movimenti rilevanti, non solo ai più evidenti).

Strutture salariali competitive nel settore pubblico che riducono il differenziale di remunerazione tra settore pubblico e privato attenuano

Δ : se il salario pubblico è comparabile a quello privato, il valore delle opportunità future nell'industria è relativamente meno determinante. Questo richiede un cambiamento nel modello di carriera del regolatore: carriere pubbliche come traiettorie valorizzate in sé, non come passi intermedi verso posizioni nell'industria.

Trasparenza dei movimenti di personale — registri pubblici delle assunzioni di ex-funzionari da parte dell'industria — crea pressione di reputazione che aumenta ϵ , il costo reputazionale associato ai comportamenti accomodanti. La trasparenza è una condizione necessaria ma non sufficiente: deve essere accompagnata da istituzioni capaci di elaborare e comunicare l'informazione pubblica.

Diversificazione delle traiettorie di carriera — promuovere movimenti del personale tra agenzie regolative, università, organizzazioni internazionali e società civile, oltre che tra regolatore e industria — riduce la concentrazione delle opportunità future nell'industria e quindi attenua l'incentivo al comportamento accomodante.

Nessuna di queste misure elimina completamente la cattura evolutiva: il vettore di pressione che la produce — la capacità dell'industria di offrire opportunità di carriera e risorse — è strutturale, non eliminabile attraverso regole individuali. Ma la combinazione delle misure può spostare significativamente l'equilibrio evolutivo p^* verso una maggiore frequenza di pratiche regolative stringenti nell'ecosistema istituzionale, rendendo la governance più resistente alla pressione selettiva dell'industria nel lungo periodo.

IV.11 — Sintesi: la mappa delle tensioni nella governance AI

Le analisi di questa Parte hanno rivelato un sistema di tensioni strutturali che caratterizzano qualsiasi tentativo di governare lo sviluppo dell'intelligenza artificiale attraverso meccanismi istituzionali. Queste tensioni non sono difetti di specifici meccanismi proposti: sono caratteristiche dello spazio dei meccanismi possibili, derivate dai fondamenti della teoria dell'informazione asimmetrica, dell'azzardo morale e della dinamica istituzionale.

La tensione informativa fondamentale: il regolatore non può sapere abbastanza. I sistemi AI sono oggetti tecnici di complessità crescente, sviluppati da organizzazioni con forti incentivi a limitare la condivisione dell'informazione tecnica sensibile. Ogni meccanismo di governance che non affronta esplicitamente questa asimmetria produce regole calibrate su informazioni distorte.

La tensione degli incentivi: le organizzazioni razionali ottimizzano rispetto alle regole. Qualsiasi requisito misurabile può diventare oggetto di ottimizzazione che massimizza il risultato sulla metrica senza soddisfare l'obiettivo sostanziale della regola. La legge di Goodhart si applica alla governance con particolare forza nei settori tecnologici avanzati, dove la distanza tra la misura e l'obiettivo può essere ampia e difficile da monitorare.

La tensione dinamica: le regole invecchiano. La frontiera tecnologica dell'AI si muove con velocità che raramente le istituzioni regolative possono seguire, creando un divario progressivo tra la realtà tecnologica

e il quadro normativo. I meccanismi di adattamento istituzionale sono più lenti del cambiamento tecnologico.

La tensione internazionale: la governance efficace richiede coordinamento in un sistema internazionale di Stati sovrani con incentivi competitivi a deviare dagli standard comuni. L'equilibrio non cooperativo produce sistematicamente livelli di sicurezza inferiori all'ottimo collettivo.

La tensione evolutiva: le istituzioni si adattano al loro ambiente. Nel lungo periodo, le pressioni selettive dell'industria tendono a erodere la stringenza regolatoria anche in assenza di corruzione o cattura intenzionale, attraverso i meccanismi della cattura cognitiva e del revolving door.

Conoscere queste tensioni non risolve il problema della governance: lo chiarisce. Una mappa accurata delle tensioni strutturali è la condizione per scegliere consapevolmente dove concentrare le risorse istituzionali limitate, quale compromesso privilegiare in ciascun contesto, e quali aspettative di performance sono realistiche per i meccanismi che è possibile costruire. La governance dell'AI non è un problema risolto — ma è un problema la cui geometria è ora sufficientemente nitida da permettere di muoversi con consapevolezza invece di procedere alla cieca.

TOPOLOGIA DEL POSSIBILE

V.1 — I limiti della matematica: cosa il formalismo rivela e cosa oscura

La potenza e il silenzio del simbolo

Ci sono cose che la matematica dice con una chiarezza impossibile al linguaggio ordinario, e cose che il linguaggio ordinario può solo sfiorare dove la matematica si arresta. L'intera architettura di questo saggio è costruita sul presupposto che entrambe le modalità di conoscenza siano necessarie e che nessuna sia sufficiente da sola. Prima di trarre le conclusioni dell'analisi, vale la pena sostare su questa distinzione — non per modestia accademica, ma perché confondere il modello con il fenomeno è il modo più comune in cui la teoria dei giochi applicata alla governance produce conclusioni fuorvianti.

La matematica impiegata nelle quattro parti precedenti rivela con precisione ineguagliabile la *geometria strutturale* dei problemi che analizza. Il teorema di Myerson-Satterthwaite non dice soltanto che è difficile progettare meccanismi di scambio efficienti con informazione privata: dice esattamente quale combinazione di proprietà è impossibile, in quale regime di parametri, e con quale forma funzionale delle tensioni che ne derivano. Questo è conoscenza di un tipo che nessuna osservazione empirica, per quanto ricca, può produrre direttamente: è conoscenza sulla struttura dello spazio dei possibili, non su un punto specifico di quello spazio.

Il formalismo rivela anche le *dipendenze logiche* tra le assunzioni e le conclusioni. Quando il VCG produce efficienza in strategie dominanti, lo fa perché i trasferimenti sono quasi-lineari, il spazio dei tipi ha certa struttura, e la funzione di allocazione massimizza il benessere dichiarato totale. Rimuovere una di queste assunzioni — tipi multidimensionali correlati, allocazioni con esternalità non catturabili nella somma delle valorizzazioni, vincoli di budget-bilanciamento — cambia la conclusione in modo determinato e tracciabile. La catena logica è trasparente, verificabile, falsificabile. Questo è precisamente il senso in cui la

matematica è superiore all'intuizione come strumento di analisi delle istituzioni.

Ma la matematica è silenziosa su domande che sono genuinamente importanti per la governance: quali obiettivi di benessere sociale meritano di essere massimizzati? Come ponderare le utilità di individui con preferenze incommensurabili? Qual è il perimetro etico della razionalità strumentale quando le preferenze degli agenti includono la preferenza a non essere manipolati? Come trattare gli interessi di generazioni future che non partecipano ai meccanismi presenti? Queste non sono domande tecniche — sono domande politiche e filosofiche che richiedono deliberazione democratica, non calcolo ottimale. Il meccanismo design può costruire il meccanismo che produce l'esito desiderato; non può dire quale esito si debba desiderare.

L'orizzonte epistemico dei modelli

I modelli formali della teoria dei giochi sono costruiti su assunzioni che li rendono trattabili: agenti con preferenze complete e transitive, tipi privati scalari in molte applicazioni, spazi di esiti finiti o compatti, distribuzioni dei tipi comuni e note, strategie scelte da insiemi convessi, assenza di esternalità non modellate. Queste assunzioni sono raramente soddisfatte esattamente nel contesto della governance AI: le preferenze dei sistemi AI non sono ancora comprese in modo sufficiente a rappresentarle come funzioni di utilità classiche; i "tipi" degli sviluppatori sono oggetti multi-dimensionalmente di alta complessità; le distribuzioni dei rischi sono profondamente incerte; le esternalità tra sistemi diversi che operano sullo stesso substrato informatico e sociale sono pervasive e difficilmente catturabili in funzioni di valorizzazione additive.

L'uso corretto del formalismo in questi contesti è il seguente: i modelli illuminano la struttura delle tensioni — mostrano quali proprietà sono in conflitto, con quale intensità, attraverso quale meccanismo causale — anche quando i parametri quantitativi sono sconosciuti o variabili. Un modello che mostra che il deficit di coordinazione cresce linearmente con il numero di attori non richiede che il numero esatto sia noto per essere informativo: suggerisce che più frammentato è il mercato degli sviluppatori, più grave è il problema di coordinazione, e questo suggerimento ha valore pratico indipendentemente dal valore preciso del coefficiente lineare.

L'abuso del formalismo consiste nel trarre conclusioni quantitative da modelli con parametri ignoti, o nel presentare i risultati di modelli con assunzioni fortemente semplificate come se fossero analisi di sistemi reali. La letteratura sulla governance AI mostra entrambi gli abusi in misura preoccupante: modelli molto semplificati calibrati su dati storici di industrie diverse vengono usati per produrre raccomandazioni di policy con falsa precisione numerica; impossibilità matematiche dimostrate in contesti specifici vengono estese in modo implicito a contesti diversi che non soddisfano le assunzioni necessarie. La consapevolezza critica di questi limiti non indebolisce la teoria — la rafforza, riducendo le sue pretese alle aree in cui la sua precisione è genuinamente giustificata.

Il problema irriducibile della rappresentazione delle preferenze

C'è un limite ancora più profondo, che tocca le fondamenta epistemiche del programma del meccanismo design. La teoria assume che le preferenze degli agenti siano *esogene al meccanismo*: il meccanismo scopre e aggrega preferenze che esistono indipendentemente da esso. Questa

assunzione è plausibile per le valorizzazioni in un'asta di frequenze radio — le preferenze degli operatori di telecomunicazione su specifiche porzioni dello spettro sono formate prima dell'asta e cambiate poco da essa. È problematica per le preferenze rilevanti alla governance AI, perché l'esistenza e il deployment diffuso di sistemi AI modifica le preferenze degli utenti — il loro modo di formare opinioni, le loro aspettative sulle interazioni sociali, la loro autonomia cognitiva — in modo che il meccanismo di governance a sua volta modifica il proprio oggetto. Le preferenze che il meccanismo dovrebbe aggregare sono parzialmente prodotte dal meccanismo stesso: un circolo che il meccanismo design standard non è equipaggiato per gestire.

Questa osservazione non invalida il meccanismo design come approccio — invalida soltanto la sua applicazione acritica senza attenzione agli effetti endogeni sulle preferenze. Il problema ha una struttura analoga a quello della causalità inversa nell'econometria: non possiamo usare lo stesso strumento per misurare la causa e l'effetto quando la causa è nel meccanismo di misurazione. Per la governance AI, significa che i meccanismi non devono essere valutati soltanto sulla base delle preferenze presenti degli agenti coinvolti, ma anche sulla base delle preferenze future che essi stessi contribuiranno a formare — una dimensione temporale e ricorsiva che nessun formalismo statico cattura pienamente.

V.2 — Le quattro tensioni strutturali della governance dell'AI

Prima tensione: efficienza versus robustezza

La prima e più fondamentale tensione strutturale è quella tra l'efficienza allocativa — massimizzare il benessere atteso dato un modello del mondo specifico — e la robustezza — mantenere prestazioni accettabili su un ampio spettro di possibili realizzazioni del mondo.

Il meccanismo bayesiano ottimale di Myerson produce il ricavo massimo quando il prior sui tipi è corretto. Se il prior è sbagliato — come accade inevitabilmente in contesti con alta incertezza sui parametri rilevanti, quale è la governance AI — il meccanismo bayesiano ottimale può performare peggio di meccanismi sub-ottimali ma robusti. Il teorema di Bergemann-Morris formalizza l'estremo robustezza: l'implementazione ex-post, equivalente all'implementazione in strategie dominanti, funziona per qualsiasi prior e qualsiasi sistema di credenze. Il suo costo è che rinuncia all'estrazione dell'informazione privata che il meccanismo bayesiano sfrutta — le rendite informative degli agenti rimangono più alte.

Questa tensione non è risolvibile in astratto: è una decisione di governance che dipende dalla fiducia del progettista nelle proprie stime dei parametri. Per la governance AI, dove i "parametri" includono le future capacità dei sistemi, le future preferenze degli utenti, e i futuri rischi sistemici — tutti oggetti radicalmente incerti — la postura prudente è privilegiare la robustezza sull'efficienza. Non perché l'efficienza non valga — è preziosissima — ma perché meccanismi efficienti sotto un modello specifico che si riveli sbagliato possono produrre esiti catastrofici, mentre meccanismi robusti producono esiti sub-ottimali ma non catastrofici in tutte le realizzazioni.

Seconda tensione: incentivi alla rivelazione versus distribuzione equa

La seconda tensione è tra il disegno degli incentivi alla rivelazione dell'informazione privata e gli obiettivi di distribuzione equa del surplus. Il meccanismo VCG produce efficienza allocativa ma non budget-bilanciamento: le rendite informative pagate agli agenti per indurre la rivelazione eccedono le tasse pagate per le esternalità. Queste rendite devono essere finanziate da una terza parte — tipicamente il contribuente — e la loro distribuzione non è necessariamente equa.

Il teorema di Myerson-Satterthwaite mostra che questa tensione è irriducibile nel contesto della contrattazione bilaterale: non esiste meccanismo che soddisfi simultaneamente efficienza, incentivi alla rivelazione, partecipazione volontaria e budget-bilanciamento. La governance AI, che richiede l'estrazione di informazione privata dagli sviluppatori sulla sicurezza dei propri sistemi, si trova inevitabilmente in questa tensione: meccanismi più efficaci nell'estrarre informazione richiedono rendite informative più alte, che devono essere finanziate (alzando il costo della governance) o accettate (rinunciando all'equità distributiva).

La scelta di quale proprietà sacrificare dipende da considerazioni di economia politica oltre che di efficienza. Un meccanismo tecnicamente ottimale che produce distribuzione del surplus percepita come ingiusta perde legittimità politica e quindi sostenibilità nel lungo periodo — un costo che la teoria non formalizza ma che è reale. Un meccanismo tecnicamente sub-ottimale che produce distribuzione percepita come equa è più robusto evolutivamente alla pressione politica e quindi più sostenibile nel tempo. L'economista politica della governance deve includere questa dimensione di legittimità che il meccanismo design puro non cattura.

Terza tensione: innovazione versus precauzione

La terza tensione è tra il ritmo dell'innovazione — che produce benefici reali rapidamente ma rischi non completamente caratterizzati — e la precauzione — che riduce i rischi ma rallenta i benefici e, in contesti competitivi, potrebbe spostare l'innovazione in giurisdizioni meno regolamentate.

Questa tensione ha una struttura di gioco multi-livello: a livello di singolo sviluppatore, il dilemma tra investire in sicurezza (riducendo la velocità del deployment) e correre (aumentando i rischi ma guadagnando vantaggio competitivo); a livello di sistema di governance nazionale, il dilemma tra stringenza (proteggendo i cittadini ma potenzialmente spostando lo sviluppo altrove) e permissività (attraendo sviluppatori e investimenti ma assumendo rischi sistemici); a livello internazionale, il dilemma tra cooperazione (producendo sicurezza globale ma cedendo sovranità tecnologica) e competizione (mantenendo autonomia tecnologica ma esacerbando i rischi sistemici).

La teoria evolutiva suggerisce che in presenza di questa struttura multi-livello di dilemmi, la soluzione non può essere trovata al solo livello micro o al solo livello macro: richiede meccanismi che coordinino i comportamenti a tutti i livelli simultaneamente, sfruttando la struttura gerarchica per creare pressioni selettive che favoriscano la cooperazione a ciascun livello. Un accordo internazionale che crea un vantaggio competitivo per i paesi che adottano standard elevati — attraverso accesso preferenziale ai mercati dei paesi partner, condivisione di risorse di ricerca, meccanismi di mutuo riconoscimento delle certificazioni — modifica i payoff del gioco tra stati in modo che trasforma la corsa al ribasso in una corsa verso l'alto.

Quarta tensione: universalità versus particolarità

La quarta tensione è tra l'aspirazione a principi universali di governance — applicabili in tutte le giurisdizioni, a tutti i sistemi, a tutti gli usi — e la necessità di particolarità contestuale: standards e processi che si adattino alle capacità specifiche dei sistemi, ai contesti specifici di uso, alle strutture istituzionali specifiche delle giurisdizioni.

I meccanismi universali sono preferibili per ragioni di comparabilità e coordinazione internazionale: permettono il mutuo riconoscimento delle certificazioni, riducono le barriere all'accesso ai mercati, facilitano la cooperazione tecnica. Ma sistemi AI con capacità molto diverse richiedono requisiti molto diversi: un sistema di raccomandazione per contenuti musicali e un sistema di supporto decisionale clinico operano in contesti di rischio incomparabili e richiedono misure di sicurezza radicalmente diverse.

La soluzione parziale è la stratificazione dei requisiti: principi generali universali (trasparenza sulle capacità, documentazione del processo di training, obbligo di disclosure degli incidenti) combinati con requisiti specifici per categoria di rischio, definiti da agenzie tecniche con mandato settoriale. Il GDPR europeo adotta questa struttura per la protezione dei dati: principi universali di trattamento, con requisiti specifici per categorie di dati speciali (sanitari, biometrici, penali). L'estensione di questa struttura alla governance AI richiede la costruzione di una tassonomia dei rischi — classificare i sistemi per categoria di rischio in modo sufficientemente granulare da produrre requisiti proporzionati, ma sufficientemente coerente da permettere la comparabilità internazionale.

V.3 — Il teorema del second-best come framework generale

L'impossibilità del primo migliore

Il teorema del second-best, enunciato da Lipsey e Lancaster nel 1956, stabilisce un principio di portata molto più ampia di quanto il suo contesto originale — la teoria del benessere e la politica commerciale — potrebbe suggerire. Il risultato è il seguente: se un sistema economico ottimizzato soddisfa un insieme di condizioni necessarie all'ottimo di Pareto, e se una di queste condizioni non può essere soddisfatta, allora la politica ottimale *non* è soddisfare tutte le altre condizioni possibili nel modo più completo possibile. Nell'ottimo del secondo migliore, può essere necessario violare molte delle condizioni di ottimo che sono soddisfacibili, perché le violazioni reciproche si compensano parzialmente.

L'intuizione è questa: un sistema con un vincolo irremovibile è un sistema diverso da quello ottimale, e ottimizzare le dimensioni libere del sistema vincolato non produce necessariamente il risultato migliore possibile. La distanza da un ottimo ideale irraggiungibile non si minimizza ottimizzando ciascuna dimensione accessibile indipendentemente dalle altre.

Applicato alla governance dell'AI, il teorema del second-best ha una conseguenza che è spesso trascurata: poiché la governance perfetta è impossibile — perché l'informazione privata degli sviluppatori non è completamente estraibile, perché l'asimmetria informativa tra regolatori e industria è strutturale, perché la coordinazione internazionale è incompleta — i meccanismi di governance ottimali del second-best

devono essere progettati come soluzioni integrate alle multiple imperfezioni del sistema, non come soluzioni locali a ciascuna imperfezione separata.

Implicazioni per il design integrato

Se non si può avere efficienza piena, incentivi perfetti alla rivelazione, distribuzione equa, e robustezza simultaneamente, la domanda non è "quale di queste proprietà scegliere" ma "quale combinazione di compromessi produce l'esito migliore nel contesto specifico?". La risposta dipende dalla struttura del problema — quali vincoli sono più stringenti, quali esternalità sono più rilevanti, qual è l'orizzonte temporale della governance — ed è diversa per la certificazione della sicurezza dei modelli fondazionali, per la governance del mercato delle applicazioni, e per il coordinamento internazionale.

Il teorema del second-best giustifica anche la cautela verso le soluzioni "semplici" nella governance AI: ogni meccanismo che risolve efficacemente uno degli aspetti del problema tende a esacerbare gli altri. La responsabilità piena risolve l'azzardo morale individuale ma non l'azzardo morale collettivo, non l'asimmetria informativa nella certificazione, e non la corsa al ribasso internazionale. I meccanismi di segnalazione credibile riducono la selezione avversa ma non garantiscono l'efficienza allocativa. Il compute governance è operativamente semplice ma non cattura i rischi emergenti da sistemi efficienti di piccola dimensione. Nessuna misura singola risolve il problema complessivo — e la tentazione di presentare misure singole come soluzioni complete è tanto più pericolosa quanto più credibile appare la singola misura.

Il second-best come postura epistemica

Oltre che come principio tecnico di design dei meccanismi, il second-best può essere adottato come postura epistemica: la riconoscenza che il mondo in cui si opera ha sempre vincoli irremovibili, e che la saggezza pratica consiste nel navigare con abilità uno spazio di possibilità ristretto invece di lamentarsi dei vincoli o fingere che non esistano.

Questa postura è l'opposto sia del realismo cinico — "le istituzioni non funzionano mai, dunque non vale la pena progettarle con cura" — sia dell'utopismo ingenuo — "se progettassimo istituzioni perfette, tutti i problemi sarebbero risolti". Il realismo rigoroso del second-best accetta i vincoli come dati, analizza la geometria dello spazio del possibile con la massima precisione disponibile, e opera con creatività all'interno di quel spazio — trovando i compromessi migliori, non fingendo che non siano necessari.

V.4 — Prospettiva d'insieme sul ciclo e sull'intelligenza collettiva

Cosa abbiamo imparato

Questo ciclo di cinque saggi ha attraversato cinque territori distinti del problema dei limiti dell'intelligenza, e il paesaggio che emerge dalla traversata è più complesso e più ricco di quanto un singolo approccio avrebbe permesso di vedere.

Il primo saggio ha mostrato che ogni sistema formale sufficientemente ricco ha proposizioni vere che non può dimostrare, e che questa

incompletezza non è una limitazione tecnica superabile con più potenza computazionale — è una struttura della logica stessa. L'intelligenza artificiale, come qualsiasi sistema computazionale, si trova inevitabilmente all'interno di questo limite: non può verificare alitmicamente tutte le proprietà rilevanti di se stesso.

Il secondo saggio ha mostrato che ogni rappresentazione comprime l'informazione, e che la geometria di Fisher-Rao è la struttura invariante che misura il costo di questa compressione nella metrica canonica delle distribuzioni di probabilità. I sistemi AI che costruiscono rappresentazioni del mondo operano sempre all'interno di questa geometria — scelgono dove perdere informazione, anche quando non lo fanno esplicitamente.

Il terzo saggio ha mostrato che i criteri di ottimizzazione non sono mai neutrali rispetto agli obiettivi che intendono promuovere: diventano obiettivi loro stessi quando vengono usati come tali, e questa distorsione — la legge di Goodhart — è strutturalmente inevitabile nell'ottimizzazione di sistemi complessi. L'allineamento degli AI è, in parte, il problema di rendere questa distorsione minimale.

Il quarto saggio ha mostrato che la verificabilità della coscienza — e più in generale delle proprietà funzionali interne degli agenti — è strutturalmente impossibile per via comportamentale: nessun insieme finito di osservazioni comportamentali può certificare le proprietà interne di un agente con capacità di ragionamento sufficientemente ricche. Questo limite investe direttamente il problema della fiducia nei sistemi AI.

Questo quinto saggio ha mostrato che l'interazione strategica tra agenti multipli produce strutture di equilibrio che nessun agente singolo

controlla e che possono divergere sistematicamente dall'ottimo collettivo: i dilemmi del prigioniero, la selezione avversa, l'azzardo morale, la cattura del regolatore, la corsa al ribasso internazionale. Questi non sono fallimenti accidentali delle istituzioni reali — sono le conseguenze strutturali dell'interazione tra razionalità individuale e benessere collettivo in presenza di informazione asimmetrica e incentivi divergenti.

L'unità profonda dei cinque territori

I cinque territori non sono separati: condividono una struttura profonda che si rivela soltanto quando vengono considerati insieme.

In ciascuno dei cinque saggi, il problema fondamentale è quello del *confine tra l'interno e l'esterno di un sistema*. In Gödel, il confine tra ciò che un sistema formale può dimostrare dall'interno e ciò che è vero senza essere dimostrabile. In Fisher-Rao, il confine tra ciò che una rappresentazione cattura e ciò che perde nella compressione. In Goodhart, il confine tra la misura e ciò che la misura intende misurare. Nel problema della coscienza, il confine tra il comportamento esterno e la vita interna. Nel meccanismo design, il confine tra l'informazione privata che un agente detiene internamente e l'informazione pubblica che il meccanismo riesce a estrarre.

In tutti e cinque i casi, l'interrogativo di fondo è: *cosa si può conoscere dall'esterno di un sistema a partire dalle sue manifestazioni esterne?* E in tutti e cinque i casi, la risposta è: meno di quanto la tentazione naturalistica suggerisca. Il limite della conoscenza dall'esterno non è tecnico — è strutturale, logico, necessario.

Questa convergenza non è accidentale. È la struttura del problema dei limiti dell'intelligenza nel senso più generale: l'intelligenza opera sempre

su rappresentazioni, le rappresentazioni sono sempre parziali, e questa parzialità non è eliminabile con più potenza computazionale, più dati, o migliori algoritmi. È iscritta nella logica della rappresentazione stessa.

L'intelligenza collettiva e il suo posto nel ciclo

La specificità del quinto territorio rispetto ai precedenti è che i limiti dell'intelligenza collettiva non sono necessariamente fissi come i limiti logici di Gödel. Le tensioni strutturali della coordinazione si attenuano con migliori istituzioni, con più informazione disponibile, con migliore progettazione dei meccanismi di incentivo. La corsa al ribasso si trasforma in corsa verso l'alto se i payoff dei giochi internazionali vengono modificati dalla struttura degli accordi. La selezione avversa si riduce se le infrastrutture di reputazione diventano più accurate e accessibili. L'azzardo morale si attenua se i meccanismi di responsabilità vengono calibrati meglio sulla struttura causale dei danni.

Questo non significa che i limiti dell'intelligenza collettiva siano eliminabili: significa che sono contingenti rispetto all'architettura istituzionale, e che questa architettura è progettabile deliberatamente, anche se mai in modo completamente ottimale. La progettazione istituzionale è il territorio dove la conoscenza tecnica degli strumenti formali si incontra con la saggezza pratica del governo — dove i teoremi della teoria dei giochi diventano, attraverso la mediazione del giudizio, principi di azione.

Il progresso dell'intelligenza collettiva è, in ultima analisi, il progresso della capacità umana di costruire istituzioni che producano esiti collettivamente desiderabili a partire da comportamenti individualmente razionali. Non è mai finito — perché le condizioni tecnologiche, economiche e politiche cambiano continuamente, richiedendo

adattamenti istituzionali continui — ma non è mai vano, perché ciascun miglioramento nella progettazione istituzionale è un guadagno reale per chi vive nelle istituzioni progettate. In questo senso, il meccanismo design è forse l'impresa intellettuale più direttamente rilevante per la condizione umana che la teoria economica abbia mai prodotto.

APPENDICI TECNICHE

Appendice A — Il teorema del punto fisso di Kakutani e l'esistenza dell'equilibrio di Nash

Il teorema di Kakutani: enunciato completo

Teorema (Kakutani, 1941): Sia X un sottoinsieme compatto e convesso di $\mathbb{R}^n$. Sia $F : X \rightarrow 2^X$ una corrispondenza con le seguenti proprietà:

- (i) $F(x)$ è non-vuota per ogni $x \in X$; (ii) $F(x)$ è convessa per ogni $x \in X$;
- (iii) il grafo di F , $\text{Gr}(F) = \{(x, y) : y \in F(x)\}$, è chiuso in $X \times X$.

Allora esiste $x^* \in X$ tale che $x^* \in F(x^*)$.

Dimostrazione via triangolazione di Sperner

Prima di applicare il teorema a Nash, presentiamo la via costruttiva alla dimostrazione che passa per il lemma di Sperner e il teorema di Brouwer, perché questa via è più elementare e illumina la struttura del risultato.

Lemma di Sperner: Sia un semplice con $n+1$ vertici triangolato in piccoli semplici, e sia assegnato a ogni vertice della triangolazione un'etichetta in $\{0, \dots, n\}$ tale che: ogni vertice del semplice originale

mantiene la propria etichetta, e ogni vertice sul lato di indice J riceve un'etichetta in J . Allora il numero di simplessi della triangolazione con tutte le $n+1$ etichette distinte è dispari (in particolare, positivo).

Dimostrazione del lemma di Sperner per induzione su n . Caso base $n=0$: triviale. Per $n=1$: considera un intervallo suddiviso in sottointervalli. L'endpoint sinistro ha etichetta 0, il destro ha etichetta 1. Ogni sottointervallo etichettato $\{0,1\}$ "completo" contribuisce ± 1 al conteggio di un certo integrale combinatorio; il risultato segue dal fatto che l'integrale è 1 perché il boundary ha segno determinato. Per n generale: si considera la struttura del boundary e si applica l'ipotesi induttiva. $\square$

Teorema di Brouwer: Ogni funzione continua $f : \Delta^n \rightarrow \Delta^n$ su un semplice ha un punto fisso.

Dimostrazione: Si approssima f con funzioni lineari a tratti su triangolazioni sempre più fini. Il lemma di Sperner garantisce l'esistenza di un semplice "completamente etichettato" in ogni triangolazione. Al limite della sequenza di triangolazioni con $\text{mesh} \rightarrow 0$, la sequenza di punti fissi approssimati ha un punto di accumulazione che è un punto fisso di f . $\square$

Da Brouwer a Kakutani: Sotto le ipotesi del teorema di Kakutani, si costruisce una sequenza di funzioni continue f_n che approssimano la corrispondenza F nel senso che $f_n(x) \in F_\varepsilon(x)$ per ogni x , dove F_ε è un allargamento ε dell'immagine di F . Per il teorema di Brouwer, ogni f_n ha un punto fisso x_n^* . Poiché X è compatto, la sequenza (x_n^*) ha un punto di accumulazione x^* . La chiusura del grafo di F garantisce $x^* \in F(x^*)$. $\square$

Applicazione all'equilibrio di Nash: verifica dettagliata delle ipotesi

Il semplice delle strategie miste $\Delta = \times_i \Delta(A_i)$ è compatto (prodotto di insiemi compatti) e convesso (combinazioni convesse di strategie miste sono strategie miste). Ipotesi di spazio verificata.

La corrispondenza di best-response congiunta $BR(\sigma) = \times_i BR_i(\sigma_{-i})$ è non-vuota per ogni σ : per ogni giocatore i , l'utilità attesa $U_i(\sigma_i, \sigma_{-i})$ è continua in σ_i su un compatto $\Delta(A_i)$, dunque il massimo è raggiunto. Ipotesi (i) verificata.

$BR_i(\sigma_{-i})$ è convessa: se σ_i e σ_i' massimizzano $U_i(\cdot, \sigma_{-i})$, e U_i è lineare in σ_i , allora ogni combinazione convessa $\lambda\sigma_i + (1-\lambda)\sigma_i'$ soddisfa:

$$\begin{aligned} U_i(\lambda\sigma_i + (1-\lambda)\sigma_i', \sigma_{-i}) &= \lambda U_i(\sigma_i, \sigma_{-i}) + (1-\lambda)U_i(\sigma_i', \sigma_{-i}) \\ &= \max_{\tau} U_i(\tau, \sigma_{-i}) \end{aligned}$$

Dunque anche la combinazione convessa è una best-response. Poiché $BR = \times_i BR_i$ e prodotto di insiemi convessi è convesso, l'ipotesi (ii) è verificata.

Il grafo di BR è chiuso: supponiamo $\sigma^k \rightarrow \sigma$ e $\tau^k \rightarrow \tau$ con $\tau^k \in BR(\sigma^k)$, cioè $U_i(\tau_i^k, \sigma_{-i}^k) \geq U_i(\varrho_i, \sigma_{-i}^k)$ per ogni ϱ_i e ogni i . Passando al limite per continuità di U_i :

$$U_i(\tau_i, \sigma_{-i}) \geq U_i(\varrho_i, \sigma_{-i}) \text{ per ogni } \varrho_i \text{ e ogni } i$$

ovvero $\tau \in BR(\sigma)$. Ipotesi (iii) verificata.

Per il teorema di Kakutani, esiste $\sigma^* \in BR(\sigma^*)$, che è per definizione un equilibrio di Nash in strategie miste. $\square$

Appendice B — Il teorema di Myerson-Satterthwaite per l'envelope theorem

Setup completo

Venditore con costo $\theta_S \sim U[c, d]$ e compratore con valorizzazione $\theta_B \sim U[a, b]$, con $[a, b] \cap [c, d] \neq \emptyset$. I meccanismi BIC con IIR hanno la forma (q, t_B, t_S) dove $q(\theta_B, \theta_S) \in [0, 1]$ è la probabilità di commercio.

Definizione: Un meccanismo è BIC (Bayesian Incentive Compatible) se la verità è best-response in attesa per ogni tipo:

$$V_B(\theta_B) = \max_{\theta'_B} \{ \theta'_B \int \theta_B Q_B(\theta'_B) - T_B(\theta'_B) \} \text{ [compratore]}$$

$$V_S(\theta_S) = \max_{\theta'_S} \{ T_S(\theta'_S) - \theta'_S \int \theta_S Q_S(\theta'_S) \} \text{ [venditore]}$$

dove $Q_B(\theta_B) = E_{\theta_S} [q(\theta_B, \theta_S)]$ e $Q_S(\theta_S) = E_{\theta_B} [q(\theta_B, \theta_S)]$.

Caratterizzazione per envelope theorem

Lemma (Payoff envelope): In ogni meccanismo BIC:

$$V_B(\theta_B) = V_B(a) + \int_a^{\theta_B} Q_B(t) dt \text{ con } Q_B \text{ non-decrescente}$$

$$V_S(\theta_S) = V_S(d) - \int_{\theta_S}^d Q_S(t) dt \text{ con } Q_S \text{ non-crescente}$$

Dimostrazione (per V_B ; V_S è analoga). Per BIC:

$$V_B(\theta_B) \geq \theta_B \int \theta_B Q_B(\theta'_B) - T_B(\theta'_B) = V_B(\theta'_B) + (\theta_B - \theta'_B) \int \theta_B Q_B(\theta'_B)$$

$$V_B(\theta_B) \geq V_B(\theta_B) + (\theta_B - \theta_B)Q_B(\theta_B)$$

Sommando le due:

$$(\theta_B - \theta_B)(Q_B(\theta_B) - Q_B(\theta_B)) \geq 0$$

Questa condizione di "complementary slackness" implica che Q_B è non-decrescente (se $\theta_B > \theta_B$ allora $Q_B(\theta_B) \geq Q_B(\theta_B)$). Dalla prima disuguaglianza divisa per $(\theta_B - \theta_B) > 0$ e passando al limite:

$$V_B'(\theta_B) = Q_B(\theta_B) \text{ q.o.}$$

$$\text{Integrando: } V_B(\theta_B) = V_B(a) + \int_a^{\theta_B} Q_B(t) dt. \square$$

Il surplus totale e la dimostrazione dell'impossibilità

Il surplus atteso totale distribuito agli agenti si calcola integrando i payoff:

$$E_{\{\theta_B\}}[V_B(\theta_B)] = V_B(a) + \int_a^b (1 - F_B(\theta_B)) Q_B(\theta_B) d\theta_B$$

[per parti]

$$= V_B(a) + \iint (1 - F_B(\theta_B))/(b-a) \cdot q(\theta_B, \theta_S) dF$$

$$E_{\{\theta_S\}}[V_S(\theta_S)] = V_S(d) + \int_c^d F_S(\theta_S) Q_S(\theta_S) d\theta_S$$

[per parti]

$$= V_S(d) + \iint F_S(\theta_S)/(d-c) \cdot q(\theta_B, \theta_S) dF$$

Somma totale del surplus agli agenti:

$$E[V_B + V_S] = V_B(a) + V_S(d) + E[\psi(\theta_B, \theta_S) \cdot q(\theta_B, \theta_S)]$$

dove $\psi(\theta_B, \theta_S) = (1-F_B(\theta_B))/f_B(\theta_B) + F_S(\theta_S)/f_S(\theta_S) > 0$
per supporti sovrapposti.

Dimostrazione dell'impossibilità. Budget-bilanciamento (BB) richiede $E[t_B + t_S] = 0$, cioè i pagamenti netti si annullano, dunque il surplus totale agli agenti uguaglia il surplus totale dello scambio:

$$E[V_B + V_S] = E[(\theta_B - \theta_S) \cdot q(\theta_B, \theta_S)]$$

Con EFF: $q(\theta_B, \theta_S) = 1_{\{\theta_B \geq \theta_S\}}$. Sostituendo:

$$V_B(a) + V_S(d) = E[(\theta_B - \theta_S) \cdot 1_{\{\theta_B \geq \theta_S\}}] - E[\psi(\theta_B, \theta_S) \cdot 1_{\{\theta_B \geq \theta_S\}}]$$

Poiché $\psi > 0$ sulla regione di intersezione $\{a \leq \theta_S \leq b, c \leq \theta_B \leq d, \theta_B \geq \theta_S\}$ — che ha misura positiva quando i supporti si sovrappongono — si ha:

$$V_B(a) + V_S(d) = E[(\theta_B - \theta_S) \cdot 1_{\{\theta_B \geq \theta_S\}}] - (\text{termine positivo}) < E[(\theta_B - \theta_S) \cdot 1_{\{\theta_B \geq \theta_S\}}]$$

Ma IIR richiede $V_B(a) \geq 0$ e $V_S(d) \geq 0$. Poiché $E[(\theta_B - \theta_S) \cdot 1_{\{\theta_B \geq \theta_S\}}] > 0$ (i supporti si sovrappongono), la disuguaglianza è consistente per i valori dei payoff individuali, ma il termine positivo deve essere strettamente positivo:

$$E[\psi \cdot 1_{\{\theta_B \geq \theta_S\}}] > 0$$

Questo termine rappresenta il surplus informativo totale — le rendite necessarie per l'incentive compatibility — che deve essere positivo. Ma esso è sottratto dal lato destro, rendendo $V_B(a) + V_S(d) < E[(\theta_B - \theta_S) \cdot 1_{\{\theta_B \geq \theta_S\}}]$. Poiché il lato sinistro deve essere ≥ 0 per IIR, e il lato destro è strettamente minore dell'aspettazione del surplus netto dello scambio efficiente, l'identità non può essere soddisfatta con EFF, BIC, IIR e BB simultaneamente. $\square$

Appendice C — La probabilità di fissazione nel processo di Moran

Il processo di Moran come catena di Markov

Consideriamo una popolazione di N individui con k cooperatori e $N-k$ defezionanti. Nel processo di Moran con intensità di selezione $w \in [0,1]$:

$$f_C(k) = 1 - w + w \cdot \pi_C(k) \quad f_D(k) = 1 - w + w \cdot \pi_D(k)$$

dove π_C e π_D sono i payoff della dinamica del gioco per cooperatori e defezionanti rispettivamente. Per il dilemma del prigioniero con matrice:

$A = \begin{bmatrix} [b-c, & -c], & [b, & 0] \end{bmatrix}$ (cooperatori-cooperatori: $b-c$;
cooperatori-defezionanti: $-c$; defezionanti-cooperatori: b ;
defezionanti-defezionanti: 0)

in popolazione well-mixed di dimensione N , con $N-1$ possibili partner per ogni individuo:

$$\pi_C(k) = [(k-1)(b-c) + (N-k)(-c)] / (N-1) = [b(k-1) - c(N-1)] / (N-1)$$

$$\pi_D(k) = [k \cdot b + (N-k-1) \cdot 0] / (N-1) = bk / (N-1)$$

La differenza:

$$\pi_D(k) - \pi_C(k) = [bk - b(k-1) + c(N-1)] / (N-1) = [b + c(N-1)] / (N-1) > 0$$

Il defezionante ottiene sempre un payoff maggiore del cooperatore, indipendentemente da k — il dilemma del prigioniero è un gioco a strategia dominante. In assenza di struttura di rete o altri meccanismi, i cooperatori vengono eliminati dalla selezione con probabilità $\rightarrow 1$ per N grande.

La probabilità di fissazione: derivazione completa

Le probabilità di transizione del processo di Moran sono:

$$T^+(k) = \Pr(k \rightarrow k+1) = [kf_C(k) / W(k)] \cdot (N-k)/N$$

$$T^-(k) = \Pr(k \rightarrow k-1) = [(N-k)f_D(k) / W(k)] \cdot k/N$$

$$T^0(k) = 1 - T^+(k) - T^-(k)$$

dove $W(k) = kf_C(k) + (N-k)f_D(k)$. Il rapporto di transizione è:

$$\gamma_k = T^-(k)/T^+(k) = f_D(k)/f_C(k)$$

La probabilità di fissazione partendo da k cooperatori è la soluzione di:

$$T^+(k)(q_{k+1} - q_k) = T^-(k)(q_k - q_{k-1})$$

con $q_0 = 0$ e $q_N = 1$.

Definendo $y_k = q_k - q_{k-1}$, la ricorrenza dà:

$$y_{k+1} = \gamma_k \cdot y_k = y_1 \cdot \prod_{j=1}^k \gamma_j$$

Dalla condizione $q_N = 1$: $\sum_{k=1}^N y_k = 1$, da cui:

$$y_1 = 1 / (1 + \sum_{k=1}^{N-1} \prod_{j=1}^k \gamma_j)$$

E quindi la probabilità di fissazione partendo da m cooperatori:

$$\varrho_m = \sum_{k=1}^m y_k = (1 + \sum_{k=1}^{m-1} \prod_{j=1}^k \gamma_j) / (1 + \sum_{k=1}^{N-1} \prod_{j=1}^k \gamma_j)$$

Il caso di selezione debole e la condizione $\varrho_1 > 1/N$

Per selezione debole ($w \rightarrow 0$), si espande al primo ordine in w :

$$\gamma_k = f_D(k)/f_C(k) = [1 - w + w\pi_D(k)] / [1 - w + w\pi_C(k)]$$

$\approx 1 + w(\pi_D(k) - \pi_C(k))$ al primo ordine

Dunque $\prod_{j=1}^k \gamma_j \approx 1 + w \sum_{j=1}^k (\pi_D(j) - \pi_C(j))$ al primo ordine.

La condizione $\varrho_1 > 1/N$ si traduce, al primo ordine in w :

$$\sum_{k=1}^{N-1} \prod_{j=1}^k \gamma_j < N - 1$$

$$\text{ovvero: } \sum_{k=1}^{N-1} \sum_{j=1}^k (\pi_D(j) - \pi_C(j)) < 0$$

cioè la *media pesata* delle differenze di payoff (pesata dal numero di stati che contribuiscono) deve essere negativa — i cooperatori devono avere payoff superiore ai defezionanti in media su tutti gli stati interni $k \in \{1, \dots, N-1\}$. Questa condizione definisce esattamente quando la cooperazione è evolutivamente favorita dalla selezione nel processo di Moran.

Appendice D — Le dinamiche di replica come gradiente naturale di Fisher: dimostrazione completa

La metrica di Fisher sulla varietà delle distribuzioni categoriche

Sia $p = (p_1, \dots, p_n) \in \Delta^{n-1} \subset \mathbb{R}^n$ il vettore delle frequenze di strategia, con Δ^{n-1} il simpleso standard aperto. La distribuzione categorica sulla stringa $s \in \{1, \dots, n\}$ è $p(s) = p_s$.

La matrice di informazione di Fisher è:

$$\begin{aligned} G(p)_{ij} &= E_s [(\partial \log p(s) / \partial p_i) (\partial \log p(s) / \partial p_j)] \\ &= \sum_s p_s \cdot (\delta_{si} / p_i) (\delta_{sj} / p_j) \\ &= \delta_{ij} / p_i \end{aligned}$$

Dunque $G(p)$ è la matrice diagonale con $G_{ii} = 1/p_i$. La metrica di Fisher sullo spazio delle distribuzioni categoriche è il tensore metrico di questa matrice: $ds^2 = \sum_i dp_i^2 / p_i$.

Questa metrica è la metrica iperbolica dello spazio delle probabilità: assegna curvatura maggiore alle regioni vicino ai bordi del simpleso (dove $p_i \rightarrow 0$), riflettendo il fatto che differenze tra distribuzioni con probabilità molto piccole su certi esiti sono più significative in termini di informazione.

L'inversa della matrice di Fisher sul simpleso

Poiché p è vincolato al simpleso ($\sum p_i = 1$), la metrica di Fisher non è invertibile direttamente: la matrice G è singolare sul vincolo. Occorre la *pseudoinversa di Moore-Penrose* ristretta allo spazio tangente del simpleso $T_p \Delta^{n-1} = \{v \in \mathbb{R}^n : \sum v_i = 0\}$.

La pseudoinversa di $G(p)$ ristretta al simpleso è:

$$G^+(p)\{ij\} = p_{-i}(\delta_{ij}) - p_{-j}$$

Verifica: $G(p) \cdot G^+(p)$ è il proiettore ortogonale su $T_p \Delta^{n-1}$.

Calcoliamo:

$$[G(p) \cdot G^+(p)]\{ij\} = \sum_k G\{ik\} \cdot G^+\{kj\} = (1/p_{-i}) \cdot p_{-i}(\delta_{ij}) - p_{-j} = \delta_{ij} - p_{-j}$$

Questo è il proiettore $P = I - p1^T$ che proietta su $T_p \Delta^{n-1}$ (sottrarre la media ponderata da ogni componente). $\square$

Dimostrazione dell'equivalenza con le dinamiche di replica

Il payoff medio è $f(p) = p^T A p = \sum_{i,j} A_{ij} p_{-i} p_{-j}$.

Il gradiente euclideo in $\mathbb{R}^n$ è:

$$(\nabla f)_{-i} = \partial f / \partial p_{-i} = 2(Ap)_{-i} = 2f_{-i}(p)$$

Il gradiente naturale ristretto al simpleso è:

$$\nabla_{-i} f = [G^+(p) \cdot \nabla f]_i = \sum_j G^+\{ij\} \cdot 2f_{-j}$$

$$= \sum_j p_{-i}(\delta_{ij} - p_{-j}) \cdot 2f_{-j}$$

$$= 2p_{-i} \sum_j (\delta_{ij} - p_{-j}) f_{-j}$$

$$= 2p_{-i} [f_{-i} - \sum_j p_{-j} f_{-j}]$$

$$= 2p_{-i} [f_{-i} - f]$$

$$= 2 \dot{p}_{-i}$$

Le dinamiche di replica $\dot{p}_{-i} = p_{-i}(f_{-i} - f)$ coincidono con la metà del gradiente naturale di f nella metrica di Fisher ristretta al simpleso. $\square$

Invarianza per reparametrizzazione e significato geometrico

Il gradiente naturale è invariante per trasformazioni diffeomorfe $\varphi : \Delta \rightarrow \Delta'$: se si parametrizzano le distribuzioni con un diverso set di parametri naturali $\theta = \varphi(p)$, il gradiente naturale di $f(\theta)$ nella metrica di Fisher indotta su Θ è lo stesso oggetto geometrico del gradiente naturale in p . Questo significa che le dinamiche di replica sono intrinsecamente definite sulla varietà delle distribuzioni — non dipendono dalla parametrizzazione scelta — a differenza del gradient ascent euclideo, che dipende arbitrariamente dalla parametrizzazione.

Il significato geometrico è il seguente: le dinamiche di replica muovono le frequenze di strategia nella direzione in cui la funzione fitness cresce più rapidamente quando ci si muove nella geometria naturale dello spazio delle distribuzioni di probabilità. Non è la direzione di crescita massimale nelle coordinate euclidee — che darebbe un bias artificiale verso le strategie con alta frequenza — ma la direzione di crescita massimale corretta per la geometria intrinseca dello spazio. La correzione per la curvatura dello spazio è esattamente il fattore p_i nella formula $\dot{p}_i = p_i(f_i - \bar{f})$: le strategie rare (p_i piccolo) cambiano frequenza più lentamente in termini assoluti, il che è corretto geometricamente perché piccoli incrementi assoluti in regioni di bassa probabilità corrispondono a grandi movimenti nella metrica di Fisher.

Appendice E — La struttura computazionale del politopo degli equilibri correlati

Gli equilibri correlati come politopo

L'insieme degli equilibri correlati di un gioco finito è determinato da un sistema di disuguaglianze lineari. Per ogni giocatore i , ogni azione $a_i \in A_i$ nel supporto di μ e ogni azione alternativa $a_i' \neq a_i$, la condizione di equilibrio correlato richiede che seguire la raccomandazione sia ottimale:

$$\sum_{a_{-i}} \mu(a_i, a_{-i}) \cdot u_i(a_i, a_{-i}) \geq \sum_{a_{-i}} \mu(a_i, a_{-i}) \cdot u_i(a_i', a_{-i})$$

Riscrivendo:

$$\sum_{a_{-i}} \mu(a_i, a_{-i}) \cdot [u_i(a_i, a_{-i}) - u_i(a_i', a_{-i})] \geq 0$$

per ogni i , ogni a_i , ogni $a_i' \neq a_i$. Insieme con i vincoli $\mu(a) \geq 0$ per ogni a e $\sum_a \mu(a) = 1$, questo definisce il politopo CE come intersezione di $\Pi_i |A_i| (|A_i| - 1)$ semispazi con il simpleso standard $\Delta(A)$.

Il politopo CE è dunque un politopo convesso compatto in $\mathbb{R}^{|A|}$. Ogni sua faccia corrisponde a un sistema di vincoli attivi (alcune condizioni di equilibrio soddisfatte come uguaglianze), e i vertici del politopo sono gli equilibri correlati "estremi" — quelli che non sono combinazioni convesse di altri equilibri correlati.

La programmazione lineare sugli equilibri correlati

La struttura di politopo rende trivialmente risolvibile in tempo polinomiale qualsiasi problema di ottimizzazione di una funzione obiettivo lineare sull'insieme degli equilibri correlati:

$$\max_{\{\mu \in \text{CE}\}} c^T \mu$$

dove $c \in \mathbb{R}^{|A|}$ è il vettore di coefficienti dell'obiettivo. Questo include, come caso speciale, la massimizzazione del benessere sociale:

$$c_a = \sum_i u_i(a) \Rightarrow c^T \mu = E_\mu[\sum_i u_i] = \text{benessere sociale atteso}$$

La soluzione si ottiene attraverso algoritmi di programmazione lineare (simplex o metodi di punto interno) in tempo polinomiale nella dimensione del problema. Il problema ha al più $|A| = \prod_i |A_i|$ variabili e $\prod_i |A_i|(|A_i|-1) + |A| + 1$ vincoli — dimensioni polinomiali in $|A|$, a fronte dell'esponenziale numero di sottogame nel gioco esteso.

Contrasto con gli equilibri di Nash: Il calcolo degli equilibri di Nash è PPAD-completo (Daskalakis, Goldberg, Papadimitriou, 2006): non è noto alcun algoritmo polinomiale, e la PPAD-hardness fornisce evidenza teorica forte che nessuno esista. Il contrasto è netto e significativo: gli equilibri correlati ottimizzabili in tempo polinomiale sono strettamente più facili da computare degli equilibri di Nash, e consentono la massimizzazione del benessere sociale su di essi.

L'algoritmo di regret-minimization e la convergenza agli equilibri correlati

Un secondo risultato computazionale di grande rilievo è che gli equilibri correlati possono essere approssimati attraverso algoritmi di *no-regret learning*: algoritmi online che, quando eseguiti parallelamente da tutti i giocatori, producono una sequenza di distribuzione empirica delle strategie che converge al politopo CE.

Un algoritmo è *no-regret* se il rimpianto cumulativo soddisfa:

$$\text{Regret}_T = \max_{a_i \in A_i} \sum_{t=1}^T [u_i(a_i, a_{-i}^t) - u_i(a_i^t, a_{-i}^t)] = o(T)$$

Il *regret externalizzato* (External Regret) è il rimpianto per non aver giocato sempre la strategia fissa migliore ex-post. L'**algoritmo Hedge** di Freund e Schapire (1997) garantisce regret esterno $O(\sqrt{T \log |A_i|})$.

Teorema (Hart e Mas-Colell, 2000): Se tutti i giocatori usano algoritmi no-regret con regret esterno $o(T)$, la distribuzione empirica delle azioni giocate converge all'insieme degli equilibri correlati del gioco.

La dimostrazione sfrutta la caratterizzazione degli equilibri correlati come distribuzioni su cui nessun giocatore ha rimpianto positivo. La convergenza è nel senso della media temporale — la distribuzione empirica delle azioni fino al tempo T — e la velocità è $O(1/\sqrt{T})$ per algoritmi con regret $O(\sqrt{T})$. Questo risultato ha implicazioni dirette per l'apprendimento in sistemi multi-agente: processi di apprendimento naturali, privi di coordinazione esplicita, convergono agli equilibri correlati invece che agli equilibri di Nash, e questa convergenza è computazionalmente efficiente.

Il politopo degli equilibri correlati ha dunque un duplice vantaggio rispetto agli equilibri di Nash: è computazionalmente trattabile per l'ottimizzazione, ed è raggiungibile attraverso apprendimento distribuito senza coordinazione centrale. Questa coppia di proprietà lo rende il punto di riferimento naturale per la progettazione di meccanismi di governance in contesti dove sia l'ottimizzazione centralizzata sia la coordinazione esplicita sono difficilmente praticabili — come l'ecosistema internazionale dello sviluppo AI.

Nota bibliografica

Le analisi di questo saggio poggiano su una letteratura tecnica che si estende su sette decenni di ricerca in teoria dei giochi, meccanismo design e teoria dei sistemi adattativi. Le opere fondamentali citate nel testo, in ordine di apparizione tematica, comprendono: von Neumann e Morgenstern (1944) per la fondazione della teoria dei giochi; Nash (1950, 1951) per l'equilibrio e la sua esistenza; Kakutani (1941) per il teorema del punto fisso; Aumann (1974, 1987) per gli equilibri correlati e la loro fondazione epistemica; Shapley (1953) per il valore cooperativo; Selten (1965, 1975) per le raffinazioni; Kreps e Wilson (1982) per l'equilibrio sequenziale; Harsanyi (1967-68) per i tipi privati; Akerlof (1970) per la selezione avversa; Spence (1973) per la segnalazione; Hurwicz (1972) per il meccanismo design; Clarke (1971) e Groves (1973) per il VCG; Green e Laffont (1977) per la caratterizzazione di unicità; Myerson (1981) per l'asta ottimale e il principio di rivelazione; Myerson e Satterthwaite (1983) per l'impossibilità nella contrattazione bilaterale; Bergemann e Morris (2005) per il meccanismo design robusto; Gale e Shapley (1962) per il matching; Maynard Smith e Price (1973) per la ESS; Fisher (1930) per il teorema fondamentale della selezione; Price (1970) per l'equazione multi-livello; Ohtsuki, Hauert, Lieberman e Nowak (2006) per la condizione $b/c > k$; Nowak e Sigmund (1998) per la reciprocità indiretta; Ohtsuki e Iwasa (2004) per le Leading Eight; Kakade (2002) per il Natural Policy Gradient; Schulman et al. (2015, 2017) per TRPO e PPO; Amari (1998) per la geometria dell'informazione; Daskalakis, Goldberg e Papadimitriou (2006) per la PPAD-completezza degli equilibri di Nash; Hart e Mas-Colell (2000) per la convergenza no-regret; Lipsey e Lancaster (1956) per il teorema del second-best; Stigler (1971) per la cattura del regolatore; Carlsson e van Damme (1993) per i giochi globali.